Investigating the Consistency of Uncertainty Sampling in Deep Active Learning

Niklas Penzel[1(✉)], Christian Reimers[1], Clemens-Alexander Brust[1], and Joachim Denzler[1,2]

[1] Computer Vision Group, Friedrich Schiller University Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany
{niklas.penzel,christian.reimers,clemens-alexander.brust, joachim.denzler}@uni-jena.de
[2] Institute of Data Science, German Aerospace Center (DLR), Mälzerstraße 3, 07745 Jena, Germany

Abstract. Uncertainty sampling is a widely used active learning strategy to select unlabeled examples for annotation. However, previous work hints at weaknesses of uncertainty sampling when combined with deep learning, where the amount of data is even more significant. To investigate these problems, we analyze the properties of the latent statistical estimators of uncertainty sampling in simple scenarios. We prove that uncertainty sampling converges towards some decision boundary. Additionally, we show that it can be inconsistent, leading to incorrect estimates of the optimal latent boundary. The inconsistency depends on the latent class distribution, more specifically on the class overlap. Further, we empirically analyze the variance of the decision boundary and find that the performance of uncertainty sampling is also connected to the class regions overlap. We argue that our findings could be the first step towards explaining the poor performance of uncertainty sampling combined with deep models.

Keywords: Uncertainty sampling · Consistency · Active learning

1 Introduction

Annotating data points is a laborious and often expensive task, especially if highly trained experts are necessary, *e.g.*, in medical areas. Hence, intelligently selecting data points out of a large collection of unlabeled examples to most efficiently use human resources is a critical problem in machine learning. Active learning (AL) is one approach that tackles this problem by iteratively using a selection strategy based on a classifier trained on some initially labeled data.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-030-92659-5_10.

C. Bauckhage et al. (Eds.): DAGM GCPR 2021, LNCS 13024, pp. 159–173, 2021.
https://doi.org/10.1007/978-3-030-92659-5_10

There are many different selection strategies possible in the AL framework [9, 11,17,24,26,27,29].

One popular strategy is uncertainty sampling, which was developed to train statistical text classifiers [17]. Uncertainty sampling uses a metric to estimate the uncertainty of the model prediction and queries the examples where the classifier is most uncertain. The original metric is the confidence of the classifier's prediction, but other popular metrics include the distance to the decision boundary, *i.e.*, margin sampling [23], and the entropy of the posterior distribution [28]. Uncertainty sampling performs well when combined with classical machine learning models, *e.g.* conditional random fields [25], support vector machines [18], or decision trees [19].

However, current state-of-the-art classification models are deep architectures like convolutional neural networks (CNNs). Combining sophisticated AL with CNNs can result in only marginal improvements [4]. One possible explanation for this behavior could be the bias introduced by the sampling strategy [21].

Towards understanding this unexpected behavior, we analyze uncertainty sampling in one-dimensional scenarios and derive the usually latent estimators of the AL system. The scenarios we investigate are closely related to binary logistic regression. Further, we can interpret the softmax activation of the output layer of a CNN as multinomial logistic regression [14, p. 266] in the extracted feature space. Hence, we argue that our approach relates to the stated goal.

Our main contribution is proving that a simple active learning system using uncertainty sampling converges against some decision boundary. We do this by analyzing the statistical estimators introduced by uncertainty sampling. To the best of our knowledge, we are the first to investigate uncertainty sampling on the level of the resulting statistical estimators. We find that the consistency depends on the latent class distribution. Furthermore, our empirical analysis reveals that the performance depends highly on the overlap of the latent class regions.

After introducing the problem in Sect. 3, we state our main findings in Sect. 4, including the proof that uncertainty sampling possibly converges towards undesired decision boundaries. Furthermore, we empirically validate and extend our findings in Sect. 5.

2 Related Work

Multiple authors report poor performances when combining different AL strategies with deep models, *i.e.*, CNNs [4,21]. Chan *et al.* conduct an ablation study with self-supervised initialization, semi-supervised learning, and AL [4]. They find that the inclusion of AL only leads to marginal benefits and speculate that the pretraining and semi-supervision already subsume the advantages.

Mittal *et al.* focus their critique on the evaluation scheme generally used to assess deep AL methods [21]. They find that changes in the training procedure significantly impact the behavior of the selection strategies. Furthermore, employing data augmentations makes the AL strategies hard to distinguish but increases overall performance. They speculate that AL may introduce a bias into the distribution of the labeled examples resulting in undesired behavior.

One bias of uncertainty sampling is visualized in the work of Sener and Savarese [24]. They use t-SNE [20] to visualize the coverage of the CIFAR-10 dataset [16] when selecting examples using uncertainty sampling. A bias towards certain areas of the feature space and a lack of selected points in other regions is visible. In contrast, the approach of Sener and Savarese leads by design to a more even coverage indicating that t-SNE truly uncovers a bias of uncertainty sampling large batches. Such a bias could be advantageous and necessary to outperform randomly sampling new points, but their empirical evaluation shows that uncertainty sampling does not perform better than random sampling.

In contrast to these previous works, we are focussing on identifying the issues of uncertainty sampling by performing a detailed theoretic analysis of an AL system in simple scenarios. We theoretically derive and investigate the usually latent parameter estimators of the decision boundary.

A significant theoretic result is provided in the work of Dasgupta [6]. They prove there is no AL strategy that can outperform random sampling in all cases. In other words, there are datasets or data distributions where randomly selecting new examples is the optimal strategy.

Additionally, Dasgupta *et al.* [7] theoretically analyze the rate of convergence of a perceptron algorithm in an AL system, but they assume linear separable classes. Similarly, the analysis of the query by committee strategy by Freund *et al.* [8] also assumes that a perfect solution exists. We do not need this assumption in our analysis. Balcan *et al.* theoretically investigate the rate of convergence of the agnostic AL strategy without the assumption of an ideal solution [2]. Also related to our theoretical approach is the work of Mussmann and Liang [22]. They focus on uncertainty sampling and logistic regression and theoretically derive bounds for the data efficiency given the inverse error rate. The authors also note that the data efficiency decreases if the means of two generated normally distributed classes are moved closer together. In contrast to these works, we focus on the consistency of the decision boundary estimators instead of the data efficiency or rate of convergence.

3 Problem Setting

Here, we describe the classifier and estimators we want to analyze in Sect. 4. We start by describing a simple binary one-dimensional problem. Let us assume there is some latent mixture consisting of two components. These components define the class distributions and can be described by the density functions p_1 and p_2. We want to determine the class of a point, *i.e.*, whether it was drawn from the distribution of class 1 or class 2. Towards this goal, we want to estimate the optimal decision boundary M. Given such a decision boundary, the classifier is a simple threshold operation.

Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be sets of sample points drawn from classes 1 and 2, respectively. We estimate the decision boundary without knowledge of the latent distribution by assuming that both classes are normally distributed. To further simplify the problem, we also assume that both classes are equally likely, *i.e.*,

the mixture weights are $\frac{1}{2}$, and that both classes share the same variance σ^2. This approach is closely related to linear discriminant analysis (LDA) [14, p. 242]. LDA also assumes normal distributions but results in logistic functions describing the probability that an example is of a certain class. Another related approach is logistic regression [14, p. 250] where a logistic function is estimated using the maximum likelihood principle. Though in our simple scenario, both LDA and logistic regression result in the same decision boundary. In contrast, we directly estimate the decision boundary. Our approach enables us to study the statistical properties of the related estimators.

Under the described assumptions, the important estimators are

$$\hat{\mu}_i = \frac{1}{|\mathcal{X}_i|} \sum_{X \in \mathcal{X}_i} X, \qquad (1) \qquad\qquad \hat{M} = \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}, \qquad (2)$$

where $i \in \{1, 2\}$ denotes the class. The mean estimators $\hat{\mu}_i$ are the maximum likelihood estimators for Gaussians [3, p. 93]. A derivation for the decision boundary estimator $\hat{M}$ can be found in Appendix A. Note that the shared variance σ^2 is not needed to estimate M. Hence, we can ignore the variance of the class distributions and do not need to estimate it.

The AL system estimates the means and decision boundaries for multiple time steps $t \geq 0$, resulting in a sequence of estimators. In the beginning we start with m_0 examples per class, *i.e.*, $|\mathcal{X}_1^{(0)}| = |\mathcal{X}_2^{(0)}| = m_0$ holds. During each time step t, we select one example X_{t+1} that we annotate before estimating the next set of parameters. Therefore, in our AL system, we get the estimators.

$$\hat{\mu}_{t,i} = \frac{1}{|\mathcal{X}_i^{(t)}|} \sum_{X \in \mathcal{X}_i^{(t)}} X, \qquad (3) \qquad\qquad \hat{M}_t = \frac{\hat{\mu}_{t,1} + \hat{\mu}_{t,2}}{2}, \qquad (4)$$

in all time steps $t \geq 0$. To analyze the convergence of such a uncertainty sampling using system in Sect. 4 we additionally introduce a concrete update formula for the mean estimators:

$$\hat{\mu}_{t+1,i} = \frac{|\mathcal{X}_i^{(t)}|}{|\mathcal{X}_i^{(t)}| + 1} \hat{\mu}_{t,i} + \frac{X_{t+1}}{|\mathcal{X}_i^{(t)}| + 1}. \qquad (5)$$

A complete derivation of Eq. (5) can be found in Appendix B.

To select examples X_{t+1}, let us assume we can generate a sample according to the latent distribution, *i.e.*, we do not have a pool of finitely many unlabeled examples. This does not assume knowledge of the latent distribution but merely that we can run the process that generates examples. Sampling according to the latent distribution is known as the random baseline in AL. After querying an example, we employ experts to annotate and add it to the corresponding class set $\mathcal{X}_i^{(t+1)}$. In the following time step, we estimate updated parameters that potentially better fit the latent distribution.

To perform uncertainty sampling instead, we use an uncertainty metric to assess different examples and query the example where the classifier is most uncertain. In our scenario, we use these metrics to calculate the example where

our classifier is most uncertain. Afterward, experts label this example according to the latent class distribution at this specific point. There are three common uncertainty metrics and we show in the following that they are equivalent in our scenario.

Claim. Given the described classifier, the uncertainty metrics (i) least confidence, (ii) margin, and (iii) entropy are equivalent and generate the same point.

Proof. To validate the claim, it is enough to show that the example generated by all three metrics is the same in any given time step t.

(i) Least confidence sampling queries the sample where the prediction confidence of the classifier is minimal. In a binary problem, this confidence is at least $\frac{1}{2}$, or else we predict the other class. The point where the probability for both classes is equal is exactly the intersection of the latest estimations of our Gaussian mixture components. Hence, the point X_{t+1} where the classifier is least confident is precisely the last decision boundary estimate $\hat{M}_t$.

(ii) Margin sampling selects the point closest to a decision boundary of the classifier. Here, margin sampling chooses the best, *i.e.*, the latest estimation of the decision boundary. Hence, the new point X_{t+1} is exactly $\hat{M}_t$.

(iii) When using entropy as an uncertainty metric, we query the example that maximizes the posterior distribution's entropy. The categorical distribution that maximizes the entropy is the uniform distribution over both classes [5]. Given our classifier, this is precisely the decision boundary where both classes are equally likely. Hence, X_{t+1} is our latest decision boundary estimate $\hat{M}_t$.

□

In this one-dimensional scenario, three commonly used uncertainty metrics select the same point which allows us to simplify our investigation. Instead of looking at different metrics, we will analyze the system that queries and annotates the latest decision boundary estimate in each step, *i.e.*, $X_{t+1} = \hat{M}_t$. Note that adding the last decision boundary estimate to either of the two classes leads to future estimates of the means being skewed towards our estimations of M.

The theoretical part of our analysis focuses mainly on AL, where we have no unlabeled pool but can generate examples directly instead. Before we start to analyze the consistency and convergence of such an AL system, we want to briefly discuss the differences to a finite pool of unlabeled data points. In real-world problems, we often do not have access to the example generating system but instead a fixed number T of unlabeled examples. Let the set $\mathfrak{U}$ be the unlabeled pool with $|\mathfrak{U}| = T$. Each datapoint X from $\mathfrak{U}$ belongs either to class 1 or 2.

Let us assume that $\mathfrak{U}$ is sampled from the undistorted latent distribution defining the problem. Given the sequence $(\mathfrak{U}_t)_{t\in\{0,\dots,T\}}$ of examples still unknown in time step t, we observe that $\mathfrak{U} = \mathfrak{U}_0 \supset \mathfrak{U}_1 \supset \dots \supset \mathfrak{U}_{T-1} \supset \mathfrak{U}_T = \emptyset$ applies.

Given such a finite pool, random sampling selects an element from $\mathfrak{U}_t$ uniformly, which is equivalent to sampling from the latent distribution. In contrast, uncertainty sampling selects the example

$$X_{t+1} = \arg\min_{X \in \mathfrak{U}_t} |\hat{M}_t - X|, \tag{6}$$

closest to the latest decision boundary estimate, because the actual $\hat{M}_t$ is likely not included in $\mathfrak{U}_t$.

4 Theoretical Investigation of Uncertainty Sampling

Towards the goal of analyzing the convergence of an uncertainty sampling system, we look at Eq. (4). Given this definition of $\hat{M}_t$ we know that exactly one of the statements $\hat{\mu}_{t,1} < \hat{M}_t < \hat{\mu}_{t,2}$, $\hat{\mu}_{t,1} > \hat{M}_t > \hat{\mu}_{t,2}$, or $\hat{\mu}_{t,1} = \hat{M}_t = \hat{\mu}_{t,2}$ applies. We now analyze these cases to show that uncertainty sampling converges.

We start with the case $\hat{\mu}_{t,1} = \hat{M}_t = \hat{\mu}_{t,2}$. In this case, the AL system already converged. No matter to which class we add $\hat{M}_t$, both $\hat{\mu}_{t,1}$ and $\hat{\mu}_{t,2}$ will not change for $t \to \infty$ which follows directly from Eq. (5).

Let us now look at the other two cases. We note that in a time step t the variables $\hat{\mu}_{t,1}$ and $\hat{\mu}_{t,2}$ define a random interval. Without loss of generality let us assume $\hat{\mu}_{t,1} < \hat{M}_t < \hat{\mu}_{t,2}$ applies. In step t, $\hat{M}_t$ is annotated and used to calculate $\hat{\mu}_{t+1,1}$ and $\hat{\mu}_{t+1,2}$. Let us assume the label turns out to be one. Then $\hat{\mu}_{t+1,2}$ is equal to $\hat{\mu}_{t,2}$. In contrast, we know that $\hat{\mu}_{t+1,1} > \hat{\mu}_{t,1}$ because of Eq. (5) and $\hat{M}_t > \hat{\mu}_{t,1}$. Further, we use Eq. (4) to see that $\hat{\mu}_{t,1} < \hat{\mu}_{t+1,1} < \hat{M}_{t+1} < \hat{\mu}_{t+1,2} = \hat{\mu}_{t,2}$ holds. The consequence is $[\hat{\mu}_{t+1,1}, \hat{\mu}_{t+1,2}] \subset [\hat{\mu}_{t,1}, \hat{\mu}_{t,2}]$. We can derive the same result if $\hat{M}_t$ is added to class 2 because then $\hat{\mu}_{t+1,1} = \hat{\mu}_{t,1}$ and $\hat{\mu}_{t+1,2} < \hat{\mu}_{t,2}$.

To show that uncertainty sampling converges, we now analyze these nested intervals. Given the sequences of estimators $(\hat{\mu}_{t,1})_{t\in\mathbb{N}_0}$, $(\hat{\mu}_{t,2})_{t\in\mathbb{N}_0}$ and $(\hat{M}_t)_{t\in\mathbb{N}_0}$, we already know that $\hat{M}_t \in [\hat{\mu}_{t,1}, \hat{\mu}_{t,2}]$ and $[\hat{\mu}_{t+1,1}, \hat{\mu}_{t+1,2}] \subset [\hat{\mu}_{t,1}, \hat{\mu}_{t,2}]$. If the length of nested intervals becomes arbitrarily small, then the intersection of these nested intervals is a single number [10, p. 29]. In other words, if the length converges towards zero then $\hat{\mu}_{t,1} \leq \hat{M}_\infty \leq \hat{\mu}_{t,2}$ is true for all $t \geq 0$ and for some value $\hat{M}_\infty$. It is enough to show that the nested intervals $[\hat{\mu}_{t,1}, \hat{\mu}_{t,2}]$ become arbitrarily small to prove that uncertainty sampling converges in our scenario.

Theorem 1. *The nested intervals* $[\hat{\mu}_{t,1}, \hat{\mu}_{t,2}]$ *in our one-dimensional scenario become arbitrarily small for* $t \to \infty$.

Proof. Let ϵ_t be the length of the interval $I_t = [\hat{\mu}_{t,1}, \hat{\mu}_{t,2}]$ in time step t. It is given by $\epsilon_t = \hat{\mu}_{t,2} - \hat{\mu}_{t,1}$. To prove that these interval lengths become arbitrarily small, we must show

$$\forall \delta > 0 : \lim_{t\to\infty} \epsilon_t < \delta. \tag{7}$$

The decision boundary in step t is exactly the middle of the interval $[\hat{\mu}_{t,1}, \hat{\mu}_{t,2}]$. This fact leads to the observations

$$\hat{M}_t = \hat{\mu}_{t,1} + \frac{\epsilon_t}{2} = \hat{\mu}_{t,2} - \frac{\epsilon_t}{2}. \tag{8}$$

In time step t let there be k_t examples in class 1 additionally to the m_0 initial samples, *i.e.*, $|\mathcal{X}_1^t| = m_0 + k_t$. Consequently, we know class 2 contains $|\mathcal{X}_2^t| = m_0 + t - k_t$ examples. Furthermore, there are two possible labels for $\hat{M}_t$.

Case 1: Let $\hat{M}_t$ be labeled as class 1. By using Eq. (5), Eq. (8) and some algebraic manipulations, we get the interval boundaries of I_{t+1}:

$$\hat{\mu}_{t+1,2} = \hat{\mu}_{t,2}, \text{ and} \tag{9}$$

$$\hat{\mu}_{t+1,1} = \frac{(m_0 + k_t)\hat{\mu}_{t,1} + \hat{M}_t}{m_0 + k_t + 1} = \frac{(m_0 + k_t)\hat{\mu}_{t,1} + \hat{\mu}_{t,1} + \frac{\epsilon_t}{2}}{m_0 + k_t + 1} \tag{10}$$

$$= \hat{\mu}_{t,1} + \frac{\epsilon_t}{2(m_0 + k_t + 1)}. \tag{11}$$

Therefore, the length ϵ_{t+1} of the interval I_{t+1} is

$$\epsilon_{t+1} = \hat{\mu}_{t,2} - \hat{\mu}_{t,1} - \frac{\epsilon_t}{2(m_0 + k_t + 1)} \tag{12}$$

$$= \epsilon_t - \frac{\epsilon_t}{2(m_0 + k_t + 1)} \leq \epsilon_t - \frac{\epsilon_t}{2(m_0 + t + 1)}. \tag{13}$$

Case 2: Let $\hat{M}_t$ be labeled as class 2. By using Eq. (5), Eq. (8) and some algebraic manipulations, we get the interval boundaries of I_{t+1}:

$$\hat{\mu}_{t+1,1} = \hat{\mu}_{t,1}, \text{ and} \tag{14}$$

$$\hat{\mu}_{t+1,2} = \frac{(m_0 + t - k_t)\hat{\mu}_{t,2} + \hat{M}_t}{m_0 + t - k_t + 1} = \frac{(m_0 + t - k_t)\hat{\mu}_{t,2} + \hat{\mu}_{t,2} - \frac{\epsilon_t}{2}}{m_0 + t - k_t + 1} \tag{15}$$

$$= \hat{\mu}_{t,2} - \frac{\epsilon_t}{2(m_0 + t - k_t + 1)}. \tag{16}$$

Therefore, the length ϵ_{t+1} of the interval I_{t+1} is

$$\epsilon_{t+1} = \hat{\mu}_{t,2} - \frac{\epsilon_t}{2(m_0 + t - k_t + 1)} - \hat{\mu}_{t,1} \tag{17}$$

$$= \epsilon_t - \frac{\epsilon_t}{2(m_0 + t - k_t + 1)} \leq \epsilon_t - \frac{\epsilon_t}{2(m_0 + t + 1)}. \tag{18}$$

In both cases we derive an upper bound for the length of the interval by increasing the denominator of a negative fraction and get

$$\epsilon_{t+1} \leq \epsilon_t - \frac{\epsilon_t}{2(m_0 + t + 1)} = \underbrace{\left(1 - \frac{1}{2(m_0 + t + 1)}\right)}_{=:l_t} \epsilon_t. \tag{19}$$

We now use this property recursively until we reach time step 0 resulting in

$$\epsilon_{t+1} \leq \epsilon_0 \prod_{j=0}^{t} l_j = \epsilon_0 \exp\left(\sum_{j=0}^{t} \log(l_j)\right). \tag{20}$$

Using this bound (Eq. (20)) and recalling Eq. (7), we must now show

$$\forall \delta > 0 : \lim_{t \to \infty} \epsilon_0 \exp\left(\sum_{j=0}^{t-1} \log(l_j)\right) < \delta. \tag{21}$$

Dividing by ϵ_0, we see it is enough to show that $\sum_{j=0}^{\infty} \log(l_j)$ diverges towards $-\infty$. Towards this goal, we use a known bound [1] of the natural logarithm and derive an upper bound for $\log(l_t)$

$$\log(l_t) \leq -\frac{1}{2m_0 + 2}\left(\frac{1}{t+1}\right). \tag{22}$$

The complete derivation can be found in Appendix C.

We can use this bound for $\log(l_t)$ and some algebraic manipulations to derive an upper bound for the limit we are interested in

$$\lim_{t \to \infty} \sum_{j=0}^{t-1} \log(l_j) \leq \lim_{t \to \infty} \sum_{j=0}^{t-1} -\frac{1}{2m_0 + 2}\left(\frac{1}{j+1}\right) = \lim_{t \to \infty} -\frac{1}{2m_0 + 2} \sum_{j=1}^{t} \frac{1}{j} \tag{23}$$

$$= -\frac{1}{2m_0 + 2} \sum_{j=1}^{\infty} \frac{1}{j}. \tag{24}$$

The harmonic series multiplied by a negative constant is an upper bound for the limit of the series of $\log(l_t)$. This upper bound diverges towards $-\infty$ because the harmonic series itself diverges towards ∞ [15]. Consequently, we know

$$\lim_{t \to \infty} \sum_{j=0}^{t-1} \log(l_j) = -\infty. \tag{25}$$

Using this limit as well as the facts $\delta > 0$ and $\epsilon_0 > 0$, we see

$$\forall \delta > 0 : \lim_{t \to \infty} \exp\left(\sum_{j=0}^{t-1} \log(l_j)\right) = 0 < \frac{\delta}{\epsilon_0}. \tag{26}$$

Hence, the interval lengths ϵ_t converge towards zero for $t \to \infty$. □

Until now, our analysis is independent of the specific latent class distribution. We have shown that in the one-dimensional case estimating the decision boundary M of a mixture of two classes for a growing number of examples using uncertainty sampling as a selection strategy converges towards some value $\hat{M}_\infty$. The question about the consistency of uncertainty sampling now reduces to: Is $\hat{M}_\infty$ equal to the optimal decision boundary M?

Let us assume, for example, that the latent distribution is a mixture of Gaussians. As already observed, $\hat{\mu}_{t,1} < \hat{M}_t < \hat{\mu}_{t,2}$ and $\hat{\mu}_{t+1,1} < \hat{M}_{t+1} < \hat{\mu}_{t+1,2}$ apply. Also either $\hat{\mu}_{t+1,1} > \hat{\mu}_{t,1}$ or $\hat{\mu}_{t+1,2} < \hat{\mu}_{t,2}$ holds true. if for any time step t,

$\hat{\mu}_{t,1}$ becomes larger than M, then the decision boundary M is not reachable anymore. In this case, $\hat{M}_\infty > M$ applies. The decision boundary M is likewise unreachable, if for any time step t, $\hat{\mu}_{t,2} < M$ applies. This behavior is possible because the probability density function of Gaussians is greater than zero for all possible examples $\hat{M}_t$ selected by uncertainty sampling. Hence, it is possible that $\hat{M}_t$ converges to a value $\hat{M}_\infty \neq M$. The stochastic process $\{\hat{M}_t\}_{t\in\mathbb{N}_0}$ defined by uncertainty sampling is not consistent given a mixture of Gaussians.

For further analysis, we look at the overlap ξ of the latent mixture [13]. Let the latent distribution be a mixture consisting of two components with the densities p_1 and p_2. The overlap of such a mixture is defined as

$$\xi = \int_{\mathbb{R}} \min(p_1(x), p_2(x))dx. \tag{27}$$

The behavior we determined for a mixture of two Gaussians occurs because both densities are greater than zero for all possible values. Hence, the overlap is greater than zero. More examples could lead to a wrong and unfixable decision boundary if the latent class regions overlap. We use the example of two Gaussians later on in Sect. 5.1 to empirically estimate the likelihood of such an event.

Let us now look at distributions without overlap. Without loss of generality let $\mu_1 < \mu_2$. If the latent distribution has separate class regions, *e.g.*, a uniform mixture, where the class regions are next to each other, $\hat{M}_t > M$ cannot be added to class 1. Equivalently a $\hat{M}_t < M$ can never be added to class 2 because the density of class 2 at such a point is zero. Assuming the density of the latent mixture is greater than zero for all $\hat{M}_t$, then these estimators of the decision boundary are consistent because $\forall t \in \mathbb{N}_0 : M \in [\hat{\mu}_{t,1}, \hat{\mu}_{t,2}]$.

In contrast, let us look at the case of a pool of T unlabeled examples $\mathfrak{U}$. Then the resulting sequence of intervals $[\hat{\mu}_{t,1}, \hat{\mu}_{t,2}]$ are not necessarily nested. There is the possibility that a later interval can be larger than the interval in step t if an example outside the interval is closest to the decision boundary estimate and therefore selected. Another way to think about it is to observe that for $t \to T$, uncertainty sampling tends towards random sampling because $\mathfrak{U}$ is an unbiased random sample from the latent distribution.

Let us assume our annotation budget is quite limited. We can only label T' out of a vast pool of T unlabeled examples. Under these circumstances, we claim that the undesirable properties of the infinite case of uncertainty sampling derived in this section approximately apply. In Sect. 5.2, we give empirical evidence towards this claim.

5 Empirical Investigation of Uncertainty Sampling

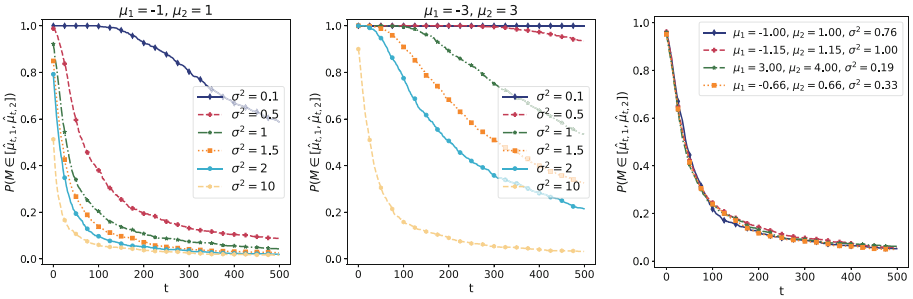
To corroborate our theoretical analysis in Sect. 4, we further investigate the consistency, convergence and performance of uncertainty sampling. In Sect. 5.1 we investigate the likelihood that the optimal decision boundary becomes unachievable. To analyze the performance we look at the empirical variance of the decision boundary estimators in Sect. 5.2.

5.1 Inconsistency Given Overlapping Class Regions

Section 4 shows that uncertainty sampling can lead to inconsistent estimators depending on the overlap of the latent mixture. Further, there are sequences of estimates where the optimal decision boundary can never be achieved. In this section, we empirically evaluate how likely such a scenario occurs. Towards this goal we want to estimate $P(M \in [\hat{\mu}_{t,1}, \hat{\mu}_{t,2}])$, *i.e.* the probability that the latent decision boundary is still achievable in step t.

To approximate these probabilities, we train an AL system 1,000 times and estimate the parameters μ_1, μ_2, and M in 500 consecutive time steps. In a given step t, the label of the estimate $\hat{M}_t$ depends on the latent mixture at this point. As a latent distribution, we select a mixture of two Gaussians with the same variance. This setup is a perfect fit for our classifier and contains an optimal latent decision boundary, *i.e.*, the intersection of both densities. To now analyze different overlapping scenarios, we repeat the experiment with different combinations of mixture parameters. We set the means either to -1 and 1, or -3 and 3, respectively. Regarding the variance, we evaluate values between 0.1 and 10.

Furthermore, to investigate if the probabilities $P(M \in [\hat{\mu}_{t,1}, \hat{\mu}_{t,2}])$ only depend on the overlap of the latent mixture, we also run the experiment for multiple parameter configurations with approximately the same overlap ξ. These configurations can be found in Table 1 in Appendix D.



(a) differing overlaps ξ (b) $\xi \approx 0.25$

Fig. 1. Estimations of the probabilities $P(M \in [\hat{\mu}_{t,1}, \hat{\mu}_{t,2}])$ for multiple parameter configurations of Gaussian mixtures. a. contains parameter compositions with differing overlap ξ. In contrast, b. displays the course of $P(M \in [\hat{\mu}_{t,1}, \hat{\mu}_{t,2}])$ for different parameter combinations defining mixtures with an overlap $\xi \approx 0.25$.

Results: Figure 1a displays the empirical estimations of the probabilities $P(M \in [\hat{\mu}_{t,1}, \hat{\mu}_{t,2}])$. We can see that the possibility that the optimal decision boundary M is achievable declines for later time steps. A smaller σ^2 and larger distances between both means, *i.e.*, more separate class regions, correlate with an increased likelihood that M is still learnable.

The influence of the overlap ξ of the latent mixture on the course of $P(M \in [\hat{\mu}_{t,1}, \hat{\mu}_{t,2}])$ can be seen in Fig. 1b. All different parameter configurations with approximately the same overlap result in roughly the same curve.

We do not know if the probabilities $P(M \in [\hat{\mu}_{t,1}, \hat{\mu}_{t,2}])$ converge towards zero for $t \to \infty$ and $\xi > 0$ but our results point in that direction. Whether this conjecture holds true or not, we find that given $\xi > 0$, uncertainty sampling is not only inconsistent, but for later time steps t, it is also unlikely that the optimal decision boundary is still achievable.

5.2 Empirical Variance of the Decision Boundary

To further analyze uncertainty sampling especially compared to random sampling, we investigate the estimators $\hat{M}_t$ using both strategies. To compare two estimators for the same parameter, we look at their respective variances. A smaller variance leads, on average, to better estimations. We approximate the variances of $\hat{M}_t$ by simulating AL systems employing uncertainty sampling or random sampling, 10,000 times with $m_0 = 3$ initial examples per class.

First, we use multiple pairs of Gaussians with approximately the same overlap ξ. Table 1 in Appendix D lists all parameter configurations.

Second, we look at two latent mixtures sharing the same class means and variance to compare separate class regions and overlapping class regions. The first one is the mixture of uniform distributions $\mathcal{U}(-2, 0)$ and $\mathcal{U}(0, 2)$. The second mixture is the mixture of Gaussians $\mathcal{N}(-1, \frac{1}{3}^2)$ and $\mathcal{N}(1, \frac{1}{3}^2)$. Derivation of these parameter configurations can be found in Appendix E.

We analyze these scenarios by first running the experiment 10,000 times, starting with $m_0 = 3$ examples per class and then estimating the empirical variance. Finally, to additionally investigate the case of a large unlabeled pool of examples, we draw $T = 25{,}006$ examples balanced from both classes in each experiment round. After three initial examples per class in $t = 0$, a pool of exactly 25,000 unlabeled examples remains. From this pool, we use Eq. (6) to select the unlabeled example closest to the estimated decision boundary in each time step. We expect similar results as in the case of AL systems directly sampling from the latent mixture.

Results. Figure 2 displays the variances of $\hat{M}_t$ for multiple time steps given latent mixtures with $\xi \approx 0.25$. We can see for all four parameter configurations that the variances converge approximately equally fast. However, they converge towards different values. The exact value seems to depend on the variance σ^2.

Figure 3 displays the results for the other empirical variance experiments. These results include the overlapping Gaussian mixtures and the separate uniform mixtures in both the finite unlabeled pool and generating examples cases. The first observation is that the results given a finite pool of unlabeled examples are nearly identical to our other results. Consequently, we conclude that our results for the infinite case of generating samples to be labeled are approximately valid for querying from a large pool of unlabeled examples.

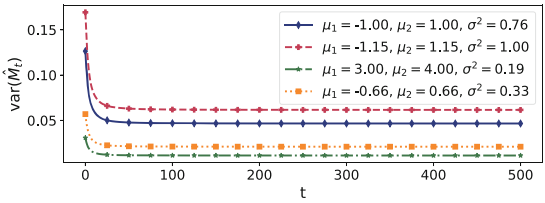


Fig. 2. The estimated empirical variance of the decision boundary estimates inbetween time steps $t = 0$ and $t = 500$. The parameter configurations define mixtures with approximately the same overlap $\xi \approx 0.25$.

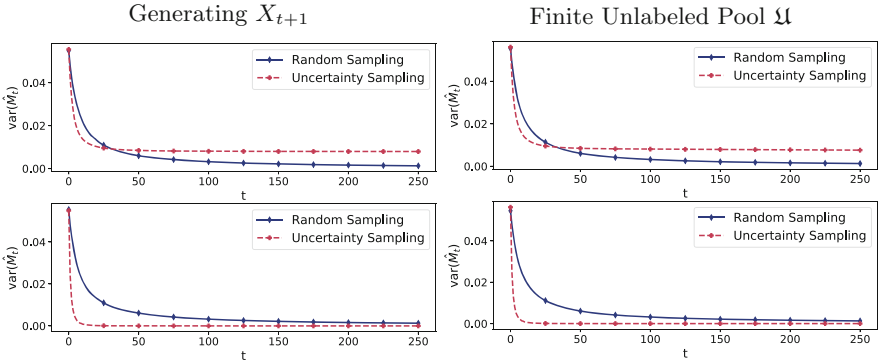


Fig. 3. The first row of plots shows the time evolution of the empirical variances of the decision boundaries in an AL scenario with overlapping class regions represented by a Gaussian mixture. The second row shows the equivalent results for a scenario with separate class regions represented by a mixture of uniform distributions. In contrast, the two columns denote generating unlabeled examples and selecting points from a finite pool, respectively.

Uncertainty sampling outperforms random sampling in the case of separate class regions. However, if the class regions overlap, uncertainty sampling performs marginally better for the first few steps but converges towards a higher, *i.e.*, worse, variance. This behavior occurs because, given overlapping class regions, the AL system converges towards different values in different experiment rounds (Sect. 5.1). Hence, given overlapping class regions, the variance of $\hat{M}_t$ is greater than zero for all time steps. In contrast, random sampling results in consistent mean estimators of the Gaussian mixture components [12, p. 217]. Hence, they stochastically converge towards μ_1 and μ_2, respectively. Consequently, the estimates of the decision boundary converge stochastically towards the optimal boundary resulting in zero variance for $t \to \infty$.

Summarizing this experiment: uncertainty sampling can outperform random sampling, but it highly depends on the latent class distribution. One reason could be the inaccurate intuition that selecting points close to the decision boundary reduces the uncertainty in this area. This intuition only holds if the class regions

of the latent distribution are separate. Given overlapping class regions, some uncertainty always remains, leading to the observed performance decrease of uncertainty sampling.

6 Conclusions

We analyzed the latent estimators of an AL system performing uncertainty sampling in one-dimensional scenarios. We find that uncertainty sampling converges, but performance and consistency depend on the overlap of the latent distribution. Our results are congruent with the work of Mussmann and Liang [22], who also find worse properties for more overlapping class regions.

Towards understanding the problems uncertainty sampling causes for deep AL systems, we reduce the classifier to the backbone feature extractor and the classification output layer. Given a CNN, we can interpret the softmax activation of the output layer as a multinomial logistic regression [14, p. 266] in the extracted feature space. Our analysis is closely related to the binary case logistic regression. Hence we argue, that this work is a first step towards understanding the underwhelming performance of uncertainty sampling combined with deep classifiers.

References

1. Upper bound of natural logarithm. https://proofwiki.org/wiki/Upper_Bound_of_Natural_Logarithm. Accessed 19 May 2021
2. Balcan, M.F., Beygelzimer, A., Langford, J.: Agnostic active learning. In: Proceedings of the 23rd International Conference on Machine Learning, ICML '06, pp. 65–72. Association for Computing Machinery, New York (2006). https://doi.org/10.1145/1143844.1143853
3. Bishop, C.M.: Pattern Recognition and Machine Learning. Information Science and Statistics, Springer, Heidelberg (2006)
4. Chan, Y.C., Li, M., Oymak, S.: On the marginal benefit of active learning: does self-supervision eat its cake? In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3455–3459. IEEE (2021)
5. Conrad, K.: Probability distributions and maximum entropy (2005)
6. Dasgupta, S.: Analysis of a greedy active learning strategy. In: Saul, L., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems, vol. 17. MIT Press (2005). https://proceedings.neurips.cc/paper/2004/file/c61fbef63df5ff317aecdc3670094472-Paper.pdf
7. Dasgupta, S., Kalai, A.T., Monteleoni, C.: Analysis of perceptron-based active learning. In: Auer, P., Meir, R. (eds.) COLT 2005. LNCS (LNAI), vol. 3559, pp. 249–263. Springer, Heidelberg (2005). https://doi.org/10.1007/11503415_17
8. Freund, Y., Seung, H.S., Shamir, E., Tishby, N.: Information, prediction, and query by committee. In: Hanson, S., Cowan, J., Giles, C. (eds.) Advances in Neural Information Processing Systems. vol. 5. Morgan-Kaufmann (1993). https://proceedings.neurips.cc/paper/1992/file/3871bd64012152bfb53fdf04b401193f-Paper.pdf

9. Freytag, A., Rodner, E., Denzler, J.: Selecting influential examples: active learning with expected model output changes. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 562–577. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_37
10. Fridy, J.A.: Introductory Analysis: The Theory of Calculus. Gulf Professional Publishing, Houston (2000)
11. Gal, Y., Islam, R., Ghahramani, Z.: Deep Bayesian active learning with image data. In: International Conference on Machine Learning, pp. 1183–1192. PMLR (2017)
12. Georgii, H.O.: Stochastik: Einführung in die Wahrscheinlichkeitstheorie und Statistik. De Gruyter (2009). publication Title: Stochastik
13. Inman, H.F., Bradley, E.L., Jr.: The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. Commun. Stat. Theory Methods **18**(10), 3851–3874 (1989)
14. Izenman, A.J.: Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer Texts in Statistics, Springer, New York (2008). https://doi.org/10.1007/978-0-387-78189-1
15. Kifowit, S.J., Stamps, T.A.: The harmonic series diverges again and again. The AMATYC Review, Spring, p. 13 (2006)
16. Krizhevsky, A.: Learning multiple layers of features from tiny images, p. 60 (2009)
17. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: Croft, B.W., van Rijsbergen, C.J. (eds.) SIGIR'94, pp. 3–12. Springer, London (1994). https://doi.org/10.1007/978-1-4471-2099-5_1
18. Loosli, G., Canu, S., Bottou, L.: Training invariant support vector machines using selective sampling. Large Scale Kernel Mach. 2 (2007)
19. Ma, L., Destercke, S., Wang, Y.: Online active learning of decision trees with evidential data. Pattern Recognit. **52**, 33–45 (2016). https://doi.org/10.1016/j.patcog.2015.10.014. https://www.sciencedirect.com/science/article/pii/S0031320315003933
20. Maaten, L.V.d., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(86), 2579–2605 (2008). http://jmlr.org/papers/v9/vandermaaten08a.html
21. Mittal, S., Tatarchenko, M., Çiçek, Ö., Brox, T.: Parting with illusions about deep active learning. arXiv preprint arXiv:1912.05361 (2019)
22. Mussmann, S., Liang, P.: On the relationship between data efficiency and error for uncertainty sampling. In: International Conference on Machine Learning, pp. 3674–3682. PMLR (2018)
23. Scheffer, T., Decomain, C., Wrobel, S.: Active hidden Markov models for information extraction. In: Hoffmann, F., Hand, D.J., Adams, N., Fisher, D., Guimaraes, G. (eds.) IDA 2001. LNCS, vol. 2189, pp. 309–318. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44816-0_31
24. Sener, O., Savarese, S.: Active learning for convolutional neural networks: a core-set approach. arXiv preprint arXiv:1708.00489 (2017)
25. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08, p. 1070. Association for Computational Linguistics (2008). https://doi.org/10.3115/1613715.1613855. http://portal.acm.org/citation.cfm?doid=1613715.1613855

26. Settles, B., Craven, M., Ray, S.: Multiple-instance active learning. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems, vol. 20, pp. 1289–1296. Curran Associates, Inc. (2008). https://proceedings.neurips.cc/paper/2007/file/a1519de5b5d44b31a01de013b9b51a80-Paper.pdf
27. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92, pp. 287–294. Association for Computing Machinery, New York (1992). https://doi.org/10.1145/130385.130417
28. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
29. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5972–5981 (2019)