ON THE EFFECTIVENESS OF ACTIVE LEARNING BY UNCERTAINTY SAMPLING IN CLASSIFICATION OF HIGH-DIMENSIONAL GAUSSIAN MIXTURE DATA

Xiaoyi Mai[⋆†]*, Salman Avestimehr*[⋆]*, Antonio Ortega*[⋆] *and Mahdi Soltanolkotabi*[⋆]

[⋆] Ming Hsieh Department of Electrical Engineering, University of Southern California, CA, USA
[†] CRIStAL, Univ. Lille, CNRS, Villeneuve d'Ascq, France.

ABSTRACT

Active learning aims to reduce the cost of labeling through selective sampling. Despite reported empirical success over passive learning, many popular active learning heuristics such as uncertainty sampling still lack satisfying theoretical guarantees. Towards closing the gap between practical use and theoretical understanding in active learning, we propose to characterize the exact behavior of uncertainty sampling for high-dimensional Gaussian mixture data, in a modern regime of big data where the numbers of samples and features are commensurately large. Through a sharp characterization of the learning results, our analysis sheds light on the important question of when uncertainty sampling works better than passive learning. Our results show that the effectiveness of uncertainty sampling is not always ensured. In fact it depends crucially on the choice of i) an adequate initial classifier used to start the active sampling process and ii) a proper loss function that allows an adaptive treatment of samples queried at various steps.

Index Terms— Active learning, uncertainty sampling, high-dimensional asymptotics, random matrix theory

1. INTRODUCTION

To achieve state of the art performance, modern machine learning techniques rely heavily on large amounts of labelled training data. However, in many application domains one has access to lots of unlabelled data and labeling the entire data set can be time consuming, expensive or both. Therefore, it is desirable to query as few labels as possible and yet achieve good accuracy. Active learning [1] aims to achieve this goal with a better selection of data to label as compared to random sampling. Despite empirical success, our theoretical understanding of popular active learning heuristics such as uncertainty sampling [2] is rather limited.

In this article we are interested in the behavior of uncertainty sampling. In uncertainty sampling, a base classifier is repeatedly trained on a growing set of labelled data, where data added at one iteration are obtained by labeling the observations with the lowest confidence level under the most recently trained base classifier. This technique of confidence-based sampling is arguably the simplest and most used active learning paradigm, with many successful empirical applications [3, 4, 5, 6, 7, 8].

Despite its popularity and its competitive performance against more sophisticated active learning methods [9, 10, 11, 12], confidence-based sampling has been analyzed in few works [13], so there is a lack of strong theoretical support for its superiority over passive learning. Meanwhile a major line of theoretical studies for active learning focuses on intractable algorithms requiring an explicit enumeration over the hypothesis space [14, 15, 16, 17]. The lack of satisfying theoretical guarantees for uncertainty sampling may be explained by the instability of confidence-sampling procedures. In fact, it has been observed that uncertainty sampling sometimes yields worse results than random sampling, depending on the learning task and the choice of hyperparameters. To capture the advantage of uncertainty sampling and how this dependence changes in different settings, a precise understanding is needed on the joint effect of data and hyperparameters.

To this aim we develop a precise characterization of confidence-based active learning in a streaming setting. Our approach draws upon recent advances in high dimensional statistics and random matrix theory to precisely predict the performance of confidence-based sampling, in a high-dimensional asymptotic regime where size of the training data and their dimension are comparably large. The high-dimensional asymptotic viewpoint provides exhaustive details into the learning performance as it varies with the sample size and the choice of hyperparameters. Compared to a series of recent works in this vein [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31], the main challenge of our analysis comes from the iterative nature of active learning, which to the best of our knowledge has not been studied before. Like the previous works, our analysis is restricted to linear models, placed under a standard setting of classifying Gaussian mixtures. As active learning involves a computationally expensive process of repeatedly training the learning model with newly added data, linear models are of particular interest in this context for their computational efficiency. Also, active learning is usually employed when labeled data are difficult to obtain, whereas the success of more complex non-linear models, e.g., neural networks, often relies on a huge amount of labeled data.

Our precise performance result, presented in Section 3, provides insights into when effective active learning can be achieved, discussed in Section 4 where we shed light on how selecting a sufficiently large initial training size, and choosing appropriate loss functions adapted to data sampled at various steps, are both critical to the success of active learning.

2. PROBLEM SETUP

Our analysis focuses on a standard Gaussian mixture model extensively studied in the setting of passive learning (see references in the introduction), with feature vectors $\mathbf{x} \in \mathbb{R}^p$ and ground truth class labels $y = \pm 1$ generated from the following distribution

$$y \sim \mathrm{Unif}\{-1, 1\}, \quad \mathbf{x} \sim \mathcal{N}(y\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{1}$$

Here, $\boldsymbol{\mu} \in \mathbb{R}^p$ are deterministic vectors of bounded norm and $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ are symmetric positive definite matrices with finite non-zero eigenvalues.

This work is supported in part by Defense Advanced Research Projects Agency (DARPA) under the Learning with Less Labels (LwLL) program.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

The Bayesian classification rule that yields the smallest error for the above Gaussian mixture model is

$$\mathbf{x}^{\mathsf{T}}\mathbf{w}^* \lessgtr 0 \Rightarrow y = \pm 1, \text{ where } \mathbf{w}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu};$$

as it is easy to check that the probability of correct classification given by any $\mathbf{w} \in \mathbb{R}^p$ satisfies the following inequality

$$\mathbb{P}(y\mathbf{w}^{\mathsf{T}}\mathbf{x} > 0|\mathbf{w}) \geq \mathrm{Err}(\mathbf{w}^*) = Q(\sqrt{u}),$$

where $Q(t) = \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-x^2/2}dx$ is the Q-function and

$$u = \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \tag{2}$$

is a quantity that reflects (inversely) the underlying difficulty of the classification problem. The goal of statistical learning is thus to estimate the oracle weight vector $\mathbf{w}^*$ from a set of labelled data.

We propose to analyze a generalized active learning procedure of confidence-based sampling, with base classifiers defined under an empirical risk minimization (ERM) framework $\min_{\mathbf{w}} \sum_i \ell(y_i\mathbf{w}^{\mathsf{T}}\mathbf{x}_i)$ which retrieves popular classification methods including SVMs, logistic regression and least-square classification with different loss functions ℓ. Since in active learning the labelled samples can no longer be seen as i.i.d. realisations due to the iterative query procedure, data queried at various iterations should be treated differently when defining the loss function. Thus, our analysis allows the use of an adaptive loss $\boldsymbol{\ell}$, defined via a set of loss functions ℓ_{tq} applied at the t-th iteration and to samples queried at the q-th iteration for some $q \leq t$. Our results apply to continuous and proper convex loss functions such as the logistic loss $\ell(s) = \ln(1 + e^{-s})$, the square loss $\ell(s) = (1-s)^2$, the exponential loss $\ell(s) = e^{-s}$, the hinge loss $\ell(t) = \max\{0, 1-t\}$ and the absolute value loss $\ell(s) = |1-s|$. The analyzed framework of active learning is summarized in Algorithm 1, where, after a initial training on randomly sampled data, the query strategy is based on the prediction classification score given by the classifier from the previous iteration, with scores of greater absolute value naturally indicating higher level of confidence in the binary classification. As uncertainty sampling typically queries the least confident samples, it corresponds to taking $\mathcal{A} = (-\epsilon, \epsilon)$ in Algorithm 1 for some small tolerance ϵ.

It is intuitively clear that a non-negligible increase in performance after a query step requires the number n_t of newly added samples not to be vanishingly small compared to the dimension p of feature vectors, in other words, the (normalized) query step sizes

$$\alpha_t = n_t/p \tag{3}$$

should be bounded away from zero in the limit of large p. More formally, this leads to the following assumption.

Assumption 1 (High-dimensional asymptotics)**.** *The initial sample ratio $\alpha_0 = n_0/p$ and the query step sizes $\alpha_t = n_t/p$ for $t \in \{0, \ldots, T\}$ are bounded away from zero for arbitrarily large p.*

Recall that $n = \sum_{t=0}^T n_t$, we define the corresponding α as

$$\alpha = n/p = \sum_{t=0}^T \alpha_t. \tag{4}$$

The empirical risk minimization involved in the training of classifier is supposed here to always be a well-posed problem with a unique solution of bounded norm. This well-posedness condition implies that α_0 is greater than 1, otherwise it is easy to show that there are infinitely many solutions for the initial classifier $\mathbf{w}_0$.

```
Algorithm 1 Confidence-Based Active Learning
Input parameters: label budget $n$, number $T$ of iterations, query strategy $\mathcal{A}$, adaptive loss $\boldsymbol{\ell} = \{\ell_{tq} | q, t \in \mathbb{N}, q \leq t, t \leq T\}$, number $n_0$ of initial training samples and sizes $n_t$ of query steps such that $\sum_{t=0}^T n_t = n$.
Obtain a initial set $\mathcal{T}$ of $n_0$ randomly selected labelled samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n_0}$, and let $\mathbf{w}_0$ be given by
    $\mathbf{w}_0 = \mathrm{argmin}_{\mathbf{w} \in \mathbb{R}^p} \sum_{i=1}^{n_0} \ell_{00}\left(y_i\mathbf{w}^{\mathsf{T}}\mathbf{x}_i\right).$
For $t \in [1, \ldots, T]$,
    1. query the label $y$ of a new coming observation $\mathbf{x}$ if $\mathbf{x}^{\mathsf{T}}\mathbf{w}_{t-1} \in \mathcal{A}$ and index the labelled sample by $(\mathbf{x}_i, y_i) \leftarrow (\mathbf{x}, y)$ with $i = |\mathcal{T}| + 1$ before adding it to the training set $\mathcal{T}$, until $|\mathcal{T}| \equiv N_q = \sum_{q=0}^t n_q$.
    2. obtain $\mathbf{w}_t$ by
        $\mathbf{w}_t = \mathrm{argmin}_{\mathbf{w} \in \mathbb{R}^p} \sum_{q=0}^t \sum_{i=N_q-n_q+1}^{N_q} \ell_{tq}\left(y_i\mathbf{w}^{\mathsf{T}}\mathbf{x}_i\right)$
Output the final active learning classifier $\mathbf{w}_T$, obtained with a total budget of $n$ label requests.
```

3. MAIN RESULTS

In this section we present our main results for predicting the active learning performance on high-dimensional Gaussian mixture data.

3.1. System of Equations

It is easy to see, by the central limit theorem, that in the limit of high dimensions, the prediction score $\mathbf{x}^{\mathsf{T}}\mathbf{w}_t$ for a new observation $\mathbf{x}$ by the classifier $\mathbf{w}_t$ of the t-th iteration is asymptotically a Gaussian variable of mean $y\mathbf{w}_t^{\mathsf{T}}\boldsymbol{\mu}$ with y the underlying class label, and variance $\mathbf{w}_t^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{w}_t$. Both $\mathbf{w}_t^{\mathsf{T}}\boldsymbol{\mu}$ and $\mathbf{w}_t^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{w}_t$can be shown to converge to some deterministic limits m_t, s_{tt}.The covariance between the prediction scores $\mathbf{x}^{\mathsf{T}}\mathbf{w}_t, \mathbf{x}^{\mathsf{T}}\mathbf{w}_{t'}$ at any two iterations has a limiting value s_{tt}. We denote by $\mathbf{m} = \{m_t\}_{t=0}^T$ the limiting mean vector of prediction scores and by $\mathbf{S} = \{s_{tt'}\}_{t,t'=0}^T$ the limiting covariance matrix.

A key fact in our characterization is that $\mathbf{m}, \mathbf{S}$ can be expressed as some deterministic functions of the asymptotic losses $\ell_{tq}(y_i\mathbf{x}_i^{\mathsf{T}}\mathbf{w}_t)$ on the training data. We then define the asymptotic loss matrix $\mathbf{C} \in \mathbb{R}^{(T+1)(T+1)}$, which is a lower triangular random matrix such that $[\mathbf{C}]_{(t+1)(q+1)}$ for $t \geq q$ has asymptotically the same distribution as $\ell_{tq}(y_i\mathbf{x}_i^{\mathsf{T}}\mathbf{w}_t)$ for $(y_i, \mathbf{x}_i)$ queried at the q-th iteration. It is far from trivial to characterize the asymptotic loss matrix $\mathbf{C}$, as the statistical behavior of the fitted scores $y_i\mathbf{x}_i^{\mathsf{T}}\mathbf{w}_t$ on the training samples $(\mathbf{x}_i, y_i)$ is hard to describe due to the implicit dependence between $\mathbf{x}_i$ and $\mathbf{w}_t$. Importantly, our analysis relies on the fact that the limiting behavior of the fitted scores can be accessed through a proximal mapping of the conditional prediction scores $y_{[q]}\mathbf{x}_{[q]}^{\mathsf{T}}\mathbf{w}_t$ for some new instance $(\mathbf{x}_{[q]}, y_{[q]})$ generated from (1) and conditioned on the query rule at the q-th iteration. Therefore our equations involve the asymptotic conditional prediction score matrix $\mathbf{R} \in \mathbb{R}^{(T+1)(T+1)}$ with $\mathbf{R}_{(t+1)(q+1)}$ following asymptotically the same distribution as $y_{[q]}\mathbf{x}_{[q]}^{\mathsf{T}}\mathbf{w}_t$. Also note that, unlike $\mathbf{C}$, $\mathbf{R}$ is statistically tractable as it contains Gaussian variables conditioned by the query rule, and its distribution can be fully described given $\mathbf{m}$ and $\mathbf{S}$: indeed, we shall define $\mathbf{R} = [\mathbf{r}_0, \ldots, \mathbf{r}_T]$ where $\mathbf{r}_0 \sim \mathcal{N}(\mathbf{m}, \mathbf{S})$ and

$$\mathbf{r}_{t+1} = \mathbf{m} + \mathbf{s}_{tt}^{-\frac{1}{2}} g_t [\mathbf{S}]_{t\cdot} + \boldsymbol{\zeta}_{t+1}, \quad t \geq 0$$

for g_t of density $f(a) = f_{\mathcal{N}(0,1)}(a) \mathbb{1}_{\mathcal{A}}(m_t + \sqrt{s_{tt}} a)$ and $\boldsymbol{\zeta}_{t+1} \sim \mathcal{N}\left(\mathbf{0}_{T+1}, \mathbf{S} - s_{tt}^{-1} [\mathbf{S}]_{\cdot t} [\mathbf{S}]_{t\cdot}\right)$ independent of g_t. The proximal operator $h_{\boldsymbol{\Phi}}(\cdot)$ that allows one to access $\mathbf{C}$ from $\mathbf{R}$ is

$$h_{\boldsymbol{\Phi}}(\mathbf{M}) = \operatorname{argmin}_{\mathbf{M}'} \Bigg[\sum_{q \leq T'} \sum_{t \geq q} \ell_{tq}([\mathbf{M}]_{(t+1)(q+1)}) + \frac{1}{2} \operatorname{tr} (\mathbf{M}' - \mathbf{M})^{\mathsf{T}} \boldsymbol{\Phi} (\mathbf{M}' - \mathbf{M}) \Bigg], \tag{5}$$

for any $\mathbf{M} \in \mathbb{R}^{(T+1)(T'+1)}$ with $T' \leq T$. With the help of the proximal operator $h_{\boldsymbol{\Phi}}(\cdot)$, the matrix $\mathbf{C}$ can be expressed as

$$\mathbf{C} = \boldsymbol{\Phi}(h_{\boldsymbol{\Phi}}(\mathbf{J} \otimes \mathbf{R}) - \mathbf{J} \otimes \mathbf{R}). \tag{6}$$

with $\mathbf{J} \in \mathbb{R}^{(T+1)(T+1)}$ and $[\mathbf{J}]_{dd'} = \mathbb{1}_{d \geq d'}$. It remains to define the deterministic matrix $\boldsymbol{\Phi}$ involved in the definition (5) of the proximal operator $h_{\boldsymbol{\Phi}}(\cdot)$ and the equations (8) and (6), which is itself the expectation of a function of $\mathbf{R}$ and $\mathbf{C}$:

$$\boldsymbol{\Phi} = -\mathbb{E}[\mathbf{C}\, \mathcal{D}(\boldsymbol{\alpha})(\mathbf{R} - \mathbf{m}\mathbf{1}_{T+1}^{\mathsf{T}})^{\mathsf{T}}] \mathbf{S}^{-1} \tag{7}$$

We are now ready to present the system of equations determining $\mathbf{m}$ and $\mathbf{S}$ as follows

$$\begin{aligned} \mathbf{S} &= \boldsymbol{\Phi}^{-1} \left(\mathbf{u} \mathbb{E}[\mathbf{C}\boldsymbol{\alpha}]^{\mathsf{T}} \mathbb{E}[\mathbf{C}\boldsymbol{\alpha}] + \mathbb{E}[\mathbf{C}\, \mathcal{D}(\boldsymbol{\alpha}) \mathbf{C}^{\mathsf{T}}] \right) (\boldsymbol{\Phi}^{-1})^{\mathsf{T}} \\ \mathbf{m} &= \boldsymbol{\Phi}^{-1} u \mathbb{E}[\mathbf{C}\boldsymbol{\alpha}] \end{aligned} \tag{8}$$

where $\boldsymbol{\alpha} = \{\alpha_t\}_{t=0}^{T}$.

3.2. Precise Performance

The primary outcome of our analysis is a prediction of the active learning performance that is exact for sufficiently high-dimensional data. As presented in Theorem 1, the asymptotic classification error can be computed by solving the system of equations (8) presented in Section 3.1. Indeed, by solving (8), we have access to the limiting value (m_t, s_{tt}) of $(\mathbf{w}_t^{\mathsf{T}} \boldsymbol{\mu}, \mathbf{w}_t^{\mathsf{T}} \boldsymbol{\Sigma} \mathbf{w}_t)$. Since the prediction score $\mathbf{w}_t^{\mathsf{T}} \mathbf{x}$ for a new observation $\mathbf{x}$ is asymptotically a Gaussian variable of mean $y m_t$ (with y the underlying class label of $\mathbf{x}$) and variance s_{tt}, we obtain the high-dimensional classification error of $\mathbf{w}_t$ as a Q-function of $m_t/\sqrt{s_{tt}}$. The proof of Theorem 1 is deferred to a longer version of this work. A numerical validation is provided in Figure 1, where a extremely close match between our theoretical prediction and the actual empirical performance is observed on data of only moderately large dimension $p = 100$.

Theorem 1 (Precise performance of Confidence-Based Active Learning)**.** *Let Assumption 1 hold. Then, for $\mathbf{w}_t$ with $t \in \{0, \ldots, T\}$ given in Algorithm 1, we have*

$$\mathbb{P}(y \mathbf{w}_t^{\mathsf{T}} \mathbf{x} > 0 | \mathbf{w}_t) = Q\left(\frac{m_t}{\sqrt{s_{tt}}}\right) + o_P(1) \tag{9}$$

with positive constants m_t, s_{tt} given in (8).

Our precise performance result allows one to assess whether active learning outperforms passive learning, leading to a series of insightful consequences discussed in Section 4

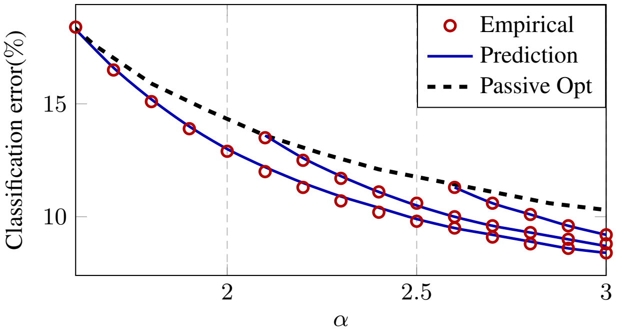

Fig. 1. Comparison of theoretical prediction (given in Theorem 1) and empirical performance (averaged over 100 realizations), along with the optimal performance of passive learning (retrieved from Theorem 2). Results reported on Gaussian mixture data with $p = 100$, $\boldsymbol{\mu} = \mathbf{1}_p/4$ and $\{\boldsymbol{\Sigma}\}_{i,j} = .4^{|i-j|}$, with $\ell_{qt} = (1-t)^2$, $\mathcal{A} = (-0.1, 0.1)$, fixed query size $n_t = 10$ and various initial numbers $n_0 = \{160, 210, 260\}$ (from left to right).

4. CONSEQUENCES

Since the goal of active learning is to surpass passive learning under the same label budget, we retrieve from [32] the best achievable passive learning performance as a point of reference for comparison.

Theorem 2 (Optimal Passive Learning Performance [32])**.** *Let Assumption 1 hold and $\mathbf{w}_{\rm ps} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^p} \sum_{i=1}^{n} \ell(y_i \mathbf{w}^{\mathsf{T}} \mathbf{x}_i)$ with n randomly sampled labelled data $(\mathbf{x}_i, y_i)$. Then,*

$$\mathbb{P}(y \mathbf{w}_{\rm ps}^{\mathsf{T}} \mathbf{x} > 0 | \mathbf{w}_{\rm ps}) \geq Q(\sqrt{\theta_{\rm ps}}) \tag{10}$$

holds with high probability for

$$\theta_{\rm ps} = \frac{(\alpha - 1) u^2}{\alpha u + 1} \tag{11}$$

where $\alpha = n/p$ per (4).

Remark also from [32] that the performance upper bound of passive learning in Theorem 2 can be reached with the square loss $\ell(t) = (1-t)^2$. We propose then to use an adaptive square loss $\boldsymbol{\ell}$ with quadratic functions ℓ_{tq} for $q \leq t$, $t \leq T$. To facilitate the discussion, we focus on one-step active learning, let $\ell_{00}(t) = \ell_{01}(t) = (1-t)^2, \ell_{11}(t) = (\lambda - t)^2$ with an adjustable $\lambda \in \mathbb{R}$ that allows an adaptive treatment of actively queried samples, and denote

$$\theta(\alpha_0, \lambda) = \lim_{\epsilon \to 0} \frac{m_1}{\sqrt{s_{11}}} \tag{12}$$

where m_1, s_{11} are given by (8) with an adaptive square loss as specified above, and with a query strategy of $\mathcal{A} = (-\epsilon, +\epsilon)$. The success of active learning is indicated by $\theta(\alpha_0, \lambda) > \theta_{\rm ps}$.

Focusing on this one-step active learning scenario, we will derive from Theorem 1 and Theorem 2 some insightful consequences on the size of initial random sampling, the optimal active learning performance and the corresponding choice of hyperparameters.

4.1. Condition on Initial Sampling

We provide first in Corollary 1 a sufficient condition on the size of initial training set so that the one-step uncertainty sampling with properly set loss function achieves a guaranteed performance gain. This result suggests that the effectiveness of uncertainty sampling can be ensured under sufficiently good initial classifiers.

Corollary 1 (Sufficient Condition on Initial Sample Size)**.** *Under Assumption 1, for any*

$$\alpha_0 > 1 + \frac{1}{u}, \tag{13}$$

we have that

$$\max_{\lambda \in \mathbb{R}} \theta(\alpha_0, \lambda) > \theta_{\text{ps}}$$

where $\theta(\alpha_0, \lambda)$ is defined by (12) *and θ_{ps} by* (11)*.*

It is often observed in practice that large-batch queries lead to performance loss. However, as revealed in Corollary 1, provided an adequate adaptive loss, the performance of active learning remains superior to that of passive learning regardless of the active query size α_1, under the condition (13) on the initial size α_0. The bound (13) on α_0 is tight with respect to the order of u, in the sense that if α_0 is smaller than $1 + \frac{1}{u}$, then there exists an α greater than α_0 such that active learning with optimal λ is surpassed by passive learning, with high probability for large n, p. Another remark to be made from (13) is that the sufficient size of initial sampling is smaller for larger u, corresponding to more separable Gaussian mixtures. This negative association between the effectiveness of active learning and the limiting error of the learning task is consistent with the observation of [13], where the authors demonstrated the link through extensive experimentation and asymptotic data efficiency.

4.2. Choice of Hyperparameters

Laid out in Corollary 2 is the optimal choice of hyperparameters α_0 and λ that maximizes the active learning performance.

Corollary 2 (Optimization of Active Learning Performance)**.** *Let Assumption 1 hold, we have*

$$\max_{\alpha_0 \in (1,\alpha), \lambda \in \mathbb{R}} \theta(\alpha_0, \lambda) = \left(1 + \frac{a\gamma^2}{(\gamma+1)^2}\right) \theta_{\text{ps}} \tag{14}$$

where $\theta(\alpha_0, \lambda)$ is defined by (12)*, θ_{ps} by* (11) *and*[1]

$$\gamma = \left(\sqrt{au - u} - 1\right)_+ .$$

Furthermore, $\theta(\alpha_0, \lambda)$ is maximized uniquely at

$$\alpha_0^* = \alpha - \frac{(\gamma+1)^2 - 1}{u} + \frac{(u+1)\gamma^2 + u\gamma}{u\gamma + u^2 + u} \tag{15}$$

$$\lambda^* = 1 + \eta \frac{(u+1)(\gamma+1)(1 - u\gamma)}{u^2 + u(\gamma+1)} \tag{16}$$

with $\eta = \frac{(a - a_0^*)au}{(a - a_0^*)(aa_0^* - 1)(u+1)u + a(a_0^* - 1)(a_0^* u + 1)} > 0.$

We remark first from (14) that given optimally set hyperparameters α_0 and λ, the minimum label budget α required for a better performance over passive learning coincides with the threshold $1 + 1/u$ of the initial sampling size α_0 to ensure an effective active learning for all values of the query step size α_1 with a properly set λ, as stated in Corollary 1. When this minimum requirement of label budget is met, the optimal value α_0^* of the initial sampling size α_0 is always greater the threshold $1 + 1/u$, which can be deduced from (15) by observing that when $\alpha > 1 + 1/u$, the first two terms at the right-hand side sum up to $1 + 1/u$ and the third term is strictly positive.

Note from (16) that the optimal value λ^* of the loss function parameter λ can be greater or smaller than 1. Since a value of λ higher than 1 can be understood as imposing class labels of larger absolute value on actively sampled data, we observe that it can be helpful to emphasize more or less the queried samples, depending on the quantities a, u, which define the learning scenario. For easier tasks with higher u and larger label budgets α, we should assign class labels of smaller absolute value to queried samples.

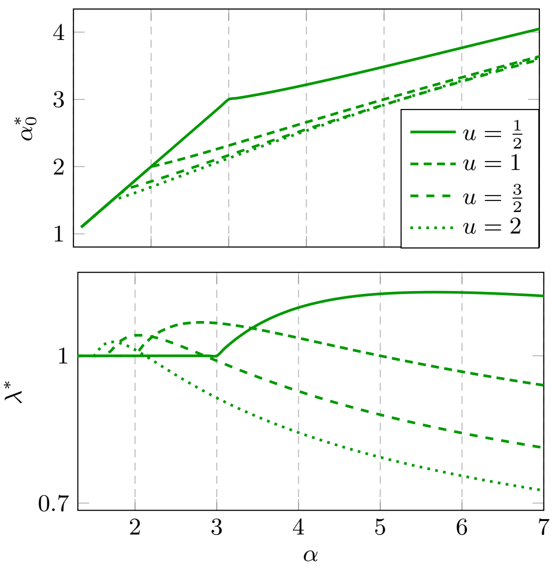

Fig. 2. The optimal choice of hyperparameters as the label budget a varies, for classification of several difficulty levels quantified by u. Top: the optimal value α_0^* (given in (15)) of the initial sampling size α_0. Bottom: the optimal value λ^* (given in (16)) of the loss function hyperparameter λ.

We visualise the behavior of optimised hyperparameters α_0^*, λ^* in Figure 2 with respect to u, a. It can be observed in the top plot of Figure 2 that after the threshold $\alpha = 1 + 1/u$, before which a totally random sampling gives the best performance as $\alpha_0^* = \alpha$, the value of α_0^* continues to grow almost linearly with α. We remark also from the bottom plot of Figure 2 that after going higher than 1 right past the threshold $\alpha = 1 + 1/u$, λ^* can descend rather quickly below 1 for a value of u as large as 2, which corresponds to an oracle classification error around 7.86%.

5. CONCLUSION

Motivated by insufficient theoretical understanding of active learning, we provided, in the limit of numerous high-dimensional data, an exact characterization of the widely used but rarely analysed method of uncertainty sampling. Our simulation showed that this high-dimensional asymptotic characterization allows one to predict the actual active learning performance on finite and moderately large data sets (of n, p around hundreds) with great precision.

A major drawback of uncertainty sampling is its instability, yielding sometimes worse performance than passive learning. As one of the first steps towards better understanding and handling the instability issue of uncertainty sampling, our study is placed under a standard data model of Gaussian mixtures which is realistic enough to demonstrate the unstable behavior of uncertainty sampling, thereby shedding light on the critical choice of hyperparameters. A more direct way to employ our results in practice is to first use prior domain knowledge to get an approximation of the oracle classification error, from which the value of u can be deduced, then plug the estimated u into our performance function to guide the choice of hyperparameters.

[1] The operator $(\cdot)_+$ preserves the input value if it is positive, and outputs zero if not.

6. REFERENCES

[1] Burr Settles, "Active learning literature survey," Tech. Rep., University of Wisconsin-Madison Department of Computer Sciences, 2009.

[2] David D Lewis and William A Gale, "A sequential algorithm for training text classifiers," in *SIGIR'94*. Springer, 1994, pp. 3–12.

[3] Hande Alemdar, TLM Van Kasteren, and Cem Ersoy, "Active learning with uncertainty sampling for large scale activity recognition in smart homes," *Journal of Ambient Intelligence and Smart Environments*, vol. 9, no. 2, pp. 209–223, 2017.

[4] Richard Segal, Ted Markowitz, and William Arnold, "Fast uncertainty sampling for labeling large e-mail corpora.," in *CEAS*. Citeseer, 2006.

[5] Greg Schohn and David Cohn, "Less is more: Active learning with support vector machines," in *ICML*. Citeseer, 2000, vol. 2, p. 6.

[6] Vikas Sindhwani, Prem Melville, and Richard D Lawrence, "Uncertainty sampling and transductive experimental design for active dual supervision," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 953–960.

[7] Alexander Liu, Goo Jun, and Joydeep Ghosh, "A self-training approach to cost sensitive uncertainty sampling," *Machine learning*, vol. 76, no. 2-3, pp. 257–270, 2009.

[8] Mustafa Bilgic, Lilyana Mihalkova, and Lise Getoor, "Active learning for networked data," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 79–86.

[9] Yazhou Yang and Marco Loog, "A benchmark and comparison of active learning for logistic regression," *Pattern Recognition*, vol. 83, pp. 401–415, 2018.

[10] Maria E Ramirez-Loaiza, Manali Sharma, Geet Kumar, and Mustafa Bilgic, "Active learning: an empirical study of common baselines," *Data mining and knowledge discovery*, vol. 31, no. 2, pp. 287–313, 2017.

[11] Burr Settles and Mark Craven, "An analysis of active learning strategies for sequence labeling tasks," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 1070–1079.

[12] Andrew I Schein and Lyle H Ungar, "Active learning for logistic regression: an evaluation," *Machine Learning*, vol. 68, no. 3, pp. 235–265, 2007.

[13] Stephen Mussmann and Percy Liang, "On the relationship between data efficiency and error for uncertainty sampling," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3674–3682.

[14] Sanjoy Dasgupta, "Coarse sample complexity bounds for active learning," *Advances in neural information processing systems*, vol. 18, pp. 235–242, 2005.

[15] Sanjoy Dasgupta, Daniel J Hsu, and Claire Monteleoni, "A general agnostic active learning algorithm," *Advances in neural information processing systems*, vol. 20, pp. 353–360, 2007.

[16] Maria-Florina Balcan, Alina Beygelzimer, and John Langford, "Agnostic active learning," *Journal of Computer and System Sciences*, vol. 75, no. 1, pp. 78–89, 2009.

[17] Chicheng Zhang and Kamalika Chaudhuri, "Beyond disagreement-based agnostic active learning," in *Advances in Neural Information Processing Systems*, 2014, pp. 442–450.

[18] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu, "On robust regression with high-dimensional predictors," *Proceedings of the National Academy of Sciences*, p. 201307842, 2013.

[19] Romain Couillet, Florent Benaych-Georges, et al., "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016.

[20] Hanwen Huang, "Asymptotic behavior of support vector machine for spiked population model," *Journal of Machine Learning Research*, vol. 18, no. 45, pp. 1–21, 2017.

[21] Edgar Dobriban, Stefan Wager, et al., "High-dimensional asymptotics of prediction: Ridge regression and classification," *The Annals of Statistics*, vol. 46, no. 1, pp. 247–279, 2018.

[22] Pragya Sur and Emmanuel J Candès, "A modern maximum-likelihood theory for high-dimensional logistic regression," *arXiv preprint arXiv:1803.06964*, 2018.

[23] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi, "Precise error analysis of regularized m-estimators in high dimensions," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5592–5628, 2018.

[24] Xiaoyi Mai and Romain Couillet, "A random matrix analysis and improvement of semi-supervised learning for large dimensional data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 3074–3100, 2018.

[25] Zhenyu Liao and Romain Couillet, "A large dimensional analysis of least squares support vector machines," *IEEE Transactions on Signal Processing*, vol. 67, no. 4, pp. 1065–1074, 2019.

[26] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi, "The impact of regularization on high-dimensional logistic regression," in *Advances in Neural Information Processing Systems*, 2019, pp. 12005–12015.

[27] Khalil Elkhalil, Abla Kammoun, Romain Couillet, Tareq Y Al-Naffouri, and Mohamed-Slim Alouini, "A large dimensional study of regularized discriminant analysis," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2464–2479, 2020.

[28] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis, "Fundamental limits of ridge-regularized empirical risk minimization in high dimensions," *arXiv preprint arXiv:2006.08917*, 2020.

[29] Christos Thrampoulidis, Samet Oymak, and Mahdi Soltanolkotabi, "Theoretical insights into multiclass classification: A high-dimensional asymptotic view," *arXiv preprint arXiv:2011.07729*, 2020.

[30] Adel Javanmard and Mahdi Soltanolkotabi, "Precise statistical analysis of classification accuracies for adversarial training," *arXiv preprint arXiv:2010.11213*, 2020.

[31] Malik Tiomoko, Romain Couillet, and Hafiz Tiomoko, "Large dimensional analysis and improvement of multi task learning," *arXiv preprint arXiv:2009.01591*, 2020.

[32] Xiaoyi Mai and Zhenyu Liao, "High Dimensional Classification via Regularized and Unregularized Empirical Risk Minimization: Precise Error and Optimal Loss," *arXiv e-prints*, p. arXiv:1905.13742, May 2019.