TOWARDS UNDERSTANDING WHEN ACTIVE LEARNING WORKS: A HIGH-DIMENSIONAL ASYMPTOTIC ANALYSIS OF UNCERTAINTY SAMPLING

Xiaoyi Mai
Department of Electrical and Computer Engineering
University of Southern California

Salman Avestimehr
Department of Electrical and Computer Engineering
University of Southern California

Antonio Ortega
Department of Electrical and Computer Engineering
University of Southern California

Mahdi Soltanolkotabi
Department of Electrical and Computer Engineering
University of Southern California

ABSTRACT

Active learning proposes to reduce the cost of labeling by selecting which data to annotate. Despite empirical superiority over passive learning, a precise theoretical understanding of widely used active learning heuristics such as uncertainty sampling remains elusive. The main challenge comes from the critical dependence of active learning heuristics on the statistical properties of the learning task and the actual practical implementation defined by a series of decisions on hyperparameters such as the base classifier, the sizes of iteration steps and the query strategy. In this paper we develop a precise analysis of active learning performance under a confidence-based sampling framework that generalizes uncertainty sampling, in a high-dimensional asymptotic regime where the dimension of feature vectors and the number of training samples are commensurately large. Our analysis allows for the derivation of precise guarantees for improvement over passive learning and actionable insights regarding the query strategy, the size of initial training set and the reliability of large-batch queries. We also utilize our framework to propose a practical scheme that adapts hyperparameter decisions to the learning task and provides a theoretically guaranteed performance gain over passive learning.

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 unlabeled 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 Settles [2009] aims to achieve this goal with a better selection of data to label as compared to random sampling. Most active learning algorithms are iterative and determine at each step, based on the information from previous iterations, the most valuable data samples to annotate and add to the training set. Despite empirical success, our theoretical understanding of popular active learning heuristics such as uncertainty sampling Lewis and Gale [1994] is rather limited.

In this article we analyze a generalized framework of uncertainty sampling (see details in Section 2.1). In uncertainty sampling, a base classifier is repeatedly trained on a growing set of labelled data, where data added at one iteration is obtained by labeling samples with 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 Alemdar et al. [2017], Segal et al. [2006], Schohn and Cohn [2000], Sindhwani et al. [2009], Liu et al. [2009], Bilgic et al. [2010].

Despite its competitive performance against more sophisticated active learning methods Yang and Loog [2018], Ramirez-Loaiza et al. [2017], Settles and Craven [2008], Schein and Ungar [2007], confidence-based sampling is rarely analyzed theoretically. The lack of theoretical analysis and corresponding guarantees may be explained by the fact that, while apparently simple, the success of confidence-sampling procedures rely critically on the learning task and the choice of hyperparameters. In fact, it has been observed that uncertainty sampling sometimes yields worse results than random sampling. To capture the advantage of confidence-based sampling and its dependence on the practical setting, a precise understanding is needed on the joint effect of data and hyperparameters. To the best of our knowledge, no such analysis has been proposed.

To this aim we develop a precise characterization of confidence-based active learning under the extensively investigated setting of linear classification on Gaussian mixtures. Our framework 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 the dimensions are comparably large. Our precise characterization provides new insights to understand the benefits and limitations of active learning over passive learning and to improve current active learning practices. We specifically discuss some of our contributions in the next section.

1.1 Contributions

Precise analysis and performance guarantee We provide the first precise performance analysis of active learning under the regime of numerous high-dimensional data, where an exact performance prediction is given for sufficiently large dimensional data. Focusing on a Gaussian mixture data model, we investigate a general algorithmic framework for confidence-based active learning, under which a base classifier is trained in an iterative fashion using empirical risk minimization, with a flexible query strategy that selects samples of a certain confidence level. Our analysis paves the way for a precise understanding on the impact of hyperparameters including the choice of base classifier, the query strategy and the sizes of query steps.

Our analysis allows us to demonstrate that with a label budget greater than a specified threshold, one can provably improve upon passive learning techniques by querying the least confident data, provided that other hyperparameters are properly set. See Theorem 3 for the formal statement.

From a technical standpoint, our analysis provides insights into the study of more sophisticated learning algorithms that are iterative in nature and apply to more complex data models. We discuss the main idea of our technical approach and lay out the key steps in Section 6.2.

Consequences and insights Our analysis reveal several consequences that both *justify and challenge* common practice, shedding light on when and how much active sampling can improve upon passive learning.

First, our results show that querying the most uncertain data does not necessarily lead to a maximized performance gain. Indeed, the relation between the confidence level and the contribution to performance depends on the data distribution and the base classifier. In particular, in some cases querying confident data can be more beneficial than uncertain data, as discussed in Section 4.1.

Our results also reveal that, if a good enough initial classifier is not available, uncertainty sampling can in fact lead to performance below that of random sampling. A condition for obtaining such good initial classifiers is provided in Theorem 4, which shows that the size of the initial training set is required to be greater than a threshold that is positively correlated with the hardness of the classification problem.

Finally, our analysis sheds light on the effect of large-batch query. Despite being useful for reducing the training cost of active learning and more convenient from the perspective of annotators' availability, large-batch query has a limited popularity due to its observed performance loss in practice when compared to random sampling. However, our results suggest that large-batch query is not an inherently bad strategy and its disadvantage with respect to random sampling can be eliminated with a properly set loss function. In fact, we find that superior performance over passive learning can always be achieved though a one-step large-batch query, as long as the batch size is greater than a threshold that depends on the underlying hardness of classification and the initial classifier (see Theorem 5).

Practical scheme for determining hyperparameters. Practical use of confidence-based sampling involves choosing several hyperparameters, such as the base classifier (e.g., classifier of SVMs, logistic regression or least-squares), batch query sizes, and query strategy (the default option being uncertainty sampling). As active learning aims to label the most informative data so that better performance can be achieved under the same label budget, cross validation cannot be employed. Indeed, in cross validation one can only judge a choice of hyperparameters by actually running the algorithm under this choice. In active learning, it means that for each candidate over a grid of hyperparameters, a different labelled set is constructed, which would require expending much more than the intended label budget.

To address the intrinsic difficulty of hyperparameter tuning in active learning, we take advantage of our analysis, which predicts the performance without actually running the active learning algorithm. The only obstacle for directly applying the performance prediction in practice is the unknown joint distribution of feature vectors and labels. Conveniently, we find that the effect of the joint distribution boils down to a scalar parameter in our analysis , which can be estimated sufficiently well from a small set of randomly selected labelled data. Based on these remarks and the previous insights, we provide in Section 5 a practical scheme for active learning with principled hyperparameter selection.

1.2 Related Work

Theoretical Analysis of Active Learning. A strong theoretical case for uncertainty sampling seems to be missing in the literature, despite its practical popularity. A major line of theoretical studies focuses on algorithms designed for an efficient search through the hypothesis space. It has been shown that this kind of active learning produces faster convergence rate than passive supervised learning in the noise-free (realizable) setting Dasgupta [2005] and is generally consistent in the sense of converging to the same solution as the passive supervised learning in the agnostic setting Dasgupta et al. [2007], Balcan et al. [2009], Zhang and Chaudhuri [2014]. Unfortunately, this approach is computationally intractable as it requires an explicit enumeration over the hypothesis space. Motivated by this limitation, several studies Huang et al. [2015], Beygelzimer et al. [2010, 2009] turned to efficient active learning algorithms that involve correcting the sampling bias caused by the non i.i.d. labelled data, and showed that these algorithms enjoy the property of consistency.

High dimensional Asymptotic Analyses. Lately a series of works have provided precise performance analyses of machine learning methods under the same high dimensional asymptotic regime as ours, ranging from purely supervised and unsupervised approaches El Karoui et al. [2013], Couillet et al. [2016], Huang [2017], Dobriban et al. [2018], Sur and Candès [2018], Thrampoulidis et al. [2018], Liao and Couillet [2019], Salehi et al. [2019], Elkhalil et al. [2020], Taheri et al. [2020], Thrampoulidis et al. [2020], Javanmard and Soltanolkotabi [2020] to techniques of semi-supervised and transfer learning Mai and Couillet [2018], Tiomoko et al. [2020]. The basis of these performance analyses is that when very high dimensional data are concerned, the performance curve varying with the amount of training data converges to a deterministic function, obtainable through concentration arguments. To the best of our knowledge, such analyses have never been conducted on active learning, which requires capturing the intricate statistical dependence among the iteratively selected labelled data.

2 Problem Setup and Preliminaries

In this section we give a formal description of the setting for our analysis, while reviewing previously established results on optimal performance of passive learning.

2.1 Framework of Confidence-Based Active Learning

A common paradigm in active learning is to train iteratively a supervised *base* classifier with active sampling at each step. In a first step, this classifier is trained with a set of randomly labeled samples. Then, at each following step, the base classifier trained on the current set is used to estimate a level of classification confidence for unlabelled items, before a confidence-based query strategy is applied to decide which new data to be labeled and added to the training set. This retraining and active learning process is repeated until the label budget is spent.

For the training of the classifier, we consider algorithms that are based on the following empirical risk minimization (ERM) framework

$$\min_{\mathbf{w}\in\mathbb{R}^p, b\in\mathbb{R}} \sum_i \ell\left(y_i(\mathbf{w}^\mathsf{T} x_i + b)\right). \tag{1}$$

Here, the weight vector $\mathbf{w}$ and and the bias term b of a linear classifier are determined by minimizing some convex loss $\ell(\cdot)$ over the training samples $(\mathbf{x}_i, y_i)$ with feature vector $\mathbf{x}_i \in \mathbb{R}^p$ and class label $y_i = \pm 1$. This ERM framework covers popular classification methods including SVMs, logistic regression and least-square classification.

To shed light on active sampling, we study a generalized query strategy that includes uncertainty sampling as a special case. This framework is summarized in Algorithm 1, where the query strategy is based on the classification score $\mathbf{w}^\mathsf{T}\mathbf{x} + b$, with scores of greater absolute value naturally indicating higher level of confidence. While a common practice is to normalize the score $\mathbf{w}^\mathsf{T}\mathbf{x} + b$ into a measure of confidence level, we skip this step in our analysis to keep the notation simple. Our framework can also handle this step with a slight modification to the approach we propose in Section 6.2. 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. We note that Algorithm 1 also allows the use of an adaptive loss $\boldsymbol{\ell}$, defined via a set of convex losses ℓ_{tq} applied at the t-th iteration and to samples queried at the q-th iteration for some $q \leq t$.

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 $\sum_{t=0}^{T} n_t = n$.
1: Obtain a initial set $\mathcal{T}$ of $n_0$ randomly selected labelled samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n_0}$, and let $\mathbf{w}_0, b_0$ be given by
       $$\min_{\mathbf{w} \in \mathbb{R}^p, b \in \mathbb{R}} \sum_{i=1}^{n_0} \ell_{00}\left(y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i + b)\right).$$
2: For $t \in [1, \ldots, T]$,
       1. query the label $y$ of a new coming observation $\mathbf{x}$ if $\mathbf{x}^\mathsf{T}\mathbf{w}_{t-1} + b_{t-1} \in \mathcal{A}$ and index the labeled 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 $\mathbf{w}_t, b_t$ by solving
              $$\min_{\mathbf{w} \in \mathbb{R}^p, b \in \mathbb{R}} \sum_{q=0}^{t} \sum_{i=N_q-n_q+1}^{N_q} \ell_{tq}\left(y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i + b)\right)$$
3: Output the final active learning classifier $(\mathbf{w}_T, b_T)$, obtained with a total budget of $n$ label requests.
```

2.2 Model and Assumptions

Our analysis focuses on the standard Gaussian mixture model extensively studied in the setting of passive learning. In this model, the feature vectors $\mathbf{x} \in \mathbb{R}^p$ and the ground truth class labels $y = \pm 1$ are generated uniformly from the following mixture model

$$y = (-1)^k \Leftrightarrow \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \quad k \in \{1, 2\}. \tag{2}$$

Here, $\boldsymbol{\mu}_k \in \mathbb{R}^p$ 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. Our results can also be generalized to non-Gaussian data as long as a certain non-sparsity condition on $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ holds with respect to the eigenspace of $\boldsymbol{\Sigma}$. We refer to the supplementary material for further details.

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

$$\mathbf{x}^\mathsf{T}\mathbf{w}_* + b_* \lessgtr 0 \Rightarrow y = \pm 1$$

with

$$\mathbf{w}_* = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \quad \text{where} \quad \boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

and $b_* = -\frac{1}{2}(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. The goal of statistical learning is thus to estimate the oracle weight vector $\mathbf{w}_*$ and bias term b_* with a limited set of labelled data. To discuss the learning performance, we define $\mathrm{Err}(\mathbf{w}, b)$ as the probability of incorrectly classifying a new observation $\mathbf{x}$ by a linear classifier $(\mathbf{w}, b)$, i.e.,

$$\mathrm{Err}(\mathbf{w}, b) = \mathbb{P}(y\mathbf{w}^\mathsf{T}\mathbf{x} + yb > 0 | \mathbf{w}, b) \tag{3}$$

with the joint distribution of $(\mathbf{x}, y)$ as described in (2). Plugging the oracle classifier $(\mathbf{w}_*, b_*)$, we have

$$\mathrm{Err}(\mathbf{w}, b) \geq \mathrm{Err}(\mathbf{w}_*, b_*) = Q(\sqrt{u})$$

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

$$u = \boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \tag{4}$$

is a quantity that reflects (inversely) the underlying difficulty of the classification problem.

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

$$\alpha_t = n_t/p \tag{5}$$

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

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

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

$$\alpha = n/p = \sum_{t=0}^{T} \alpha_t. \tag{6}$$

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 $(\mathbf{w}_0, b_0)$. Note also that that the setting of Algorithm 1 implies an unlimited source of unlabelled observations $\mathbf{x}$. Our technical approach (Section 6.2) can be easily adapted to limited sets of unlabelled data at the expense of more complicated notations.

We assume the convex loss functions ℓ_{tq} in Algorithm 1 satisfy the following regularity conditions.

Assumption 2 (Loss function)**.** *The function $\ell : \mathbb{R} \to \mathbb{R}_+$ is convex, continuous and analytic except on a finite set of points, with $\ell(t)$ finite for bounded t and $\ell'(0) < 0$.*

Assumption 2 is satisfied for many popular losses including the logistic loss $\ell(t) = \ln(1 + e^{-t})$, the square loss $\ell(t) = (1 - t)^2$, the exponential loss $\ell(t) = e^{-t}$, the hinge loss $\ell(t) = \max\{0, 1 - t\}$ and the absolute value loss $\ell(t) = |1 - t|$.

2.3 Optimal Passive Learning Performance

Since the goal of active learning is to surpass passive learning under the same label budget, we state the best achievable passive learning performance (derived in Mai and Liao [2019]) below as a point of reference for comparison with active learning.

Theorem 1 (Optimal Passive Learning Performance Mai and Liao [2019])**.** *Let Assumptions 1 and 2 hold and let $(\mathbf{w}_{\text{ps}}, b_{\text{ps}})$ stand for the solution to* (1) *with n randomly sampled labelled data $(\mathbf{x}_i, y_i)$. Then,*

$$\max_{\ell} \operatorname{Err}(\mathbf{w}_{\text{ps}}, b_{\text{ps}}) = Q(\sqrt{\theta_{\text{ps}}}) + o_P(1) \tag{7}$$

with

$$\theta_{\text{ps}} = (\alpha - 1)u^2/(\alpha u + 1) \tag{8}$$

Here, $\alpha = n/p$ per (6)*.*

3 Main Results

In this section we present our main results which predict the performance/provide guarantees for active learning.

3.1 Precise Performance Analysis

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 2, the asymptotic classification error can be computed by solving a system of equations presented later in Section 6.1. We provide in Figure 1 a comparison between our theoretical prediction and the actual empirical performance. A extremely close match is observed on data of only moderately large dimension $p = 100$.

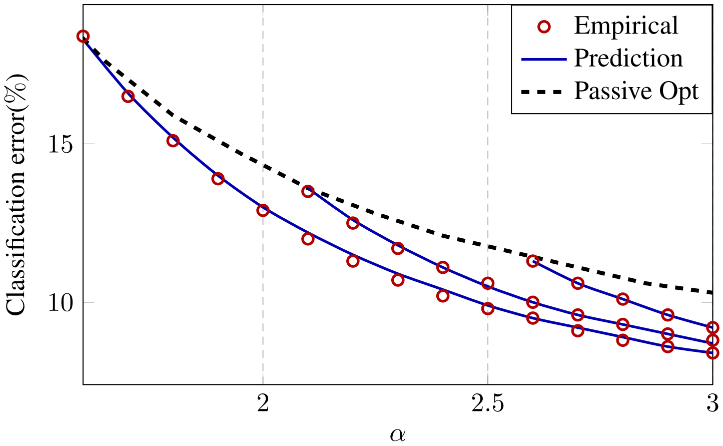

Figure 1: Comparison of theoretical prediction (given in Theorem 2) and empirical performance (averaged over 100 realizations), along with the optimal performance of passive learning (retrieved from Theorem 1). Results reported on Gaussian mixture data of $p = 100$, $\boldsymbol{\mu} = \mathbf{1}_p/4$ 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).

Theorem 2 (Precise performance of Confidence-Based Active Learning)**.** *Let Assumption 1–2 hold and* $(\mathbf{w}_T, b_T)$ *be given by Algorithm 1. Then,*

$$\operatorname{Err}(\mathbf{w}_T, b_T) = Q\left(\sqrt{\theta}\right) + o_P(1) \tag{9}$$

where

$$\theta = m_T^2/s_{TT} \tag{10}$$

with positive constants m_T, s_{TT} *given in* (17).

It is important to remark, by comparing (9) and (7), that under our high dimensional asymptotic regime active learning will outperform passive learning whenever $\theta > \theta_{\mathrm{ps}}$ for θ and θ_{ps} given in (10) and (8).

The precise performance of confidence-based active learning given in Theorem 2 is the main result of this article, allowing us to obtain guarantees for uncertainty sampling (Section 3.2) and analyze the effects of confidence level, initial classifier and large-batch query (Section 4).

3.2 Guarantees for Uncertainty Sampling

Based on the precise performance result for active learning presented in Theorem 2, we now establish theoretical guarantees for uncertainty sampling in Theorem 3 below. Essentially, Theorem 3 provides a sufficient condition on the label budget for uncertainty sampling to outperform passive learning, and a lower bound on the best achievable performance gain over passive learning. As can be seen in Theorem 3, both the label budget condition (11) and the lower bounded performance gain (12) depend on the parameter u, which reflects the separability of Gaussian mixtures. We note if separability is better, a smaller label budget is required for active learning to outperform passive learning. Alternatively, for a given label budget the advantage of active learning is greater when separability is higher.

Theorem 3 (Condition and Performance Gain of Uncertainty Sampling)**.** *Under the definitions and conditions of Theorem 2 and Theorem 1, and letting* $\mathcal{A} = (-\epsilon, \epsilon)$*, for any* $T > 1$ *and*

$$\alpha > 1 + 1/u, \tag{11}$$

we have that

$$\lim_{\epsilon \to 0} \max_{\boldsymbol{\ell}, \alpha_t} \theta > \left[1 + \alpha/(1-\gamma)^2\right] \theta_{\mathrm{ps}} \tag{12}$$

where $\gamma = 1/\sqrt{u\alpha - u}$*.*

4 Implications and Insights

In this section, we discuss how our precise analysis can be exploited to derive important consequences of practical and theoretical significance. Remark from Mai and Liao [2019] that the performance upper bound of passive learning in Theorem 1 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 on the query strategy, the initial classifier and the query sizes, we focus on one-step active learning and let $\ell_{00}(a) = \ell_{10}(a) = (1-a)^2$ and $\ell_{11}(a) = (\lambda - a)^2$ for some optimally set $\lambda \in \mathbb{R}$. Moreover, to reflect the effect of the confidence level, we denote

$$\theta_\tau = \max_{\lambda \in \mathbb{R}} \lim_{\epsilon \to 0} \theta \tag{13}$$

where θ is as defined in (10) for active classifiers obtained through one-step query with an adaptive square loss as specified above, and with a query strategy of $\mathcal{A} = (-\tau - \epsilon, -\tau + \epsilon) \cup (\tau - \epsilon, \tau + \epsilon)$ for some $\tau \in \mathbb{R}$. Recall from earlier discussions that large absolute values for scores (i.e., large τ) corresponds to higher confidence level, and the success of active learning is indicated by $\theta_\tau > \theta_{\rm ps}$.

4.1 Confidence Level and Informativeness

The question of paramount importance in active learning is which samples are the most helpful to the learning of classification boundary. Uncertainty sampling settles this question by regarding the most uncertain data, which are the closest to the current classification boundary, as the most informative ones. To investigate theoretically the presumed equality between uncertainty and informativeness, we define

$$\delta_\tau = \lim_{\alpha_1 \to 0} (\theta_{\rm ps}^{-1} - \theta_\tau^{-1})/\alpha_1$$

where δ_θ reflects the performance improvement with a incremental query step (i.e., $\alpha_1 \to 0$).

Applying Theorem 2, we have

$$\delta_\tau = -\frac{(\rho_\tau \tau - 1)^2 + (1-\rho_\tau)^2 \tau^2}{\alpha_0(\alpha_0 - 1)m_0^2} + \frac{(u+1)}{(\alpha_0 - 1)^2 u^2} \tag{14}$$

where $m_0 = \frac{u}{u+1}$ and $\rho_\tau = \frac{f_{\mathcal{N}(0,s_0)}(m_0 - \tau) - f_{\mathcal{N}(0,s_0)}(m_0 + \tau)}{f_{\mathcal{N}(0,s_0)}(m_0 - \tau) + f_{\mathcal{N}(0,s_0)}(m_0 + \tau)}$ for $f_{\mathcal{N}(0,s_0)}$ the density function of $\mathcal{N}(0, s_0)$ with $s_0 = \frac{\alpha_0 u + 1}{(\alpha_0 - 1)(u+1)^2}$. It is easy to observe from (14) that for $\rho_\tau \approx 1$, δ_θ is approximately a quadratic function of τ. Therefore, labelling data of high confidence (large τ) may sometimes lead to better performance than uncertainty sampling.

4.2 Importance of Initial Classifier

As evidenced in the expression (14) of δ_τ, querying the most uncertain data is not always guaranteed to be better than random sampling. We provide in Theorem 4 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. Note also from (15) that the threshold $1 + 1/u$ for the ratio α_0 of initial sampling is higher at smaller u, which implies, as expected, that more training samples are required on difficult tasks with less separable Gaussian mixtures.

Theorem 4 (Condition of Initial Training Set). *Under Assumptions 1 and 2, for any*

$$\alpha_0 > 1 + 1/u, \tag{15}$$

we have that

$$\theta_0 > \theta_{\rm ps}$$

where θ_0 is given by (13) *with $\tau = 0$ and $\theta_{\rm ps}$ by* (8).

4.3 Reliable Large-Batch Query

Despite the reported underperformance of large-batch query in the literature, our results suggest that the risk of large-batch query can be removed with well constructed learning models. As revealed in Theorem 4, the performance of active learning remains superior to that of passive learning regardless of the active query size α_1, under the condition (15) on the initial size α_0. Moreover, large values of α_1 are actually beneficial for achieving effective active learning beyond the condition (15), as stated in the below theorem.

Theorem 5 (Condition of Large-Batch Query)**.** *Under Assumptions 1 and 2, for any*

$$\alpha_1 > (u+1)/u^2(\alpha_0 - 1) - \alpha_0, \tag{16}$$

we have that

$$\theta_0 > \theta_{\mathrm{ps}}$$

where θ_0 is given by (13) *with $\tau = 0$ and θ_{ps} by* (8)*.*

5 Practical Scheme

As explained in the introduction, choosing the hyperparameters presents a particular challenge in active learning, where the common approach of cross validation cannot be employed. Our precise performance analysis enjoys two properties that make it applicable to the practical hyperparameter tuning: 1) it allows a sharp prediction of the active learning performance that depends on the hyperparameters and the joint distribution of data vectors and labels; 2) in our result, the effect of the joint distribution is observed to be completely characterized by *one scalar* parameter, u defined in (4), which can be well estimated with a small set of randomly labelled data.

Recall from Theorem 3 that it is possible to outperform passive learning as long as the label budget is $\alpha > 1 + 1/u$. We propose in Algorithm 2 a practical procedure that first estimates u, then checks the condition $\alpha > 1 + 1/u$ with estimated u, before optimizing the hyperparameters for a guaranteed active learning performance gain under our model of high dimensional Gaussian mixtures.

Algorithm 2 Principled Active Learning with Guarantees

```
Input parameters: label budget $n$, iteration number $T$,
sample size for estimation $\tilde{n}_0$, tolerance $\epsilon$.

Obtain a set of $\tilde{n}_0$ randomly labelled samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{\tilde{n}_0}$.
Estimate $u$ by $\hat{u} = (\kappa - p/\tilde{n}_0)/(1-\kappa)$ where $\kappa = \frac{1}{\tilde{n}_0}\sum_{i=1}^{\tilde{n}_0} y_i \mathbf{x}_i^\mathsf{T} \left(\sum_{i=1}^{\tilde{n}_0} \mathbf{x}_i \mathbf{x}_i^\mathsf{T}\right)^{-1} y_i \mathbf{x}_i$
Check if the condition $\alpha = n/p > 1 + 1/\hat{u}$ is satisfied before continuing.
Define $\hat{\theta}$ as given in (10) with estimated $\hat{u}$.
Let $\mathcal{A} = (-\epsilon, \epsilon)$, $\ell_{tq}(a) = (\lambda_{tq} - a)^2$ for all $q \leq t, t \leq T$, and find $\hat{\lambda}_{qt}, \hat{n}_t$ as solution to $\min_{\hat{\lambda}_{qt} \in \mathbb{R}, n_t \in \mathbb{N}^*} \hat{\theta}$ under
    the constraints $\sum_{t=0}^{T} n_t = n$ and $n_0 \geq \tilde{n}_0$.
Compute the active learning classifier $(\mathbf{w}_T, b_T)$ by Algorithm 1 with $\hat{n}_t$, $\ell_{tq}(a) = (\hat{\lambda}_{tq} - a)^2$, $\mathcal{A} = (-\epsilon, \epsilon)$.
```

We propose to validate our scheme in the challenging setting of large-batch query, where active learning has often been noted to underperform. The performance of one-step active learning is reported in Figure 2, where we observe an unstable performance of heuristic active learning (with non-adaptive square losses $\ell_{qt} = (1-t)^2$) that depends greatly on the size n_0 of initial set. In contrast, the principled active learning proposed in Algorithm 2 is found to yield a consistently significant performance gain over passive learning.

6 Technical Details

We begin by laying out the equations needed for Theorem 2, before giving an overview of our technical approach.

6.1 System of Equations

Under the high dimensional asymptotic regime we consider, the random variables $\mathbf{w}_t^\mathsf{T}\boldsymbol{\mu}, \mathbf{w}_t^\mathsf{T}\boldsymbol{\Sigma}\mathbf{w}_{t'}$ can be shown to converge to deterministic limits $m_t, s_{tt'}$, where we recall $\mathbf{w}_t$ is the weight vector obtained at the t-th iteration.

We define $\mathbf{m} = \{m_t\}_{t=0}^T$ and $\mathbf{S} = \{s_{tt'}\}_{t,t'=0}^T$, which are determined by the following fixed-point equations:

$$\begin{aligned} \mathbf{S} &= \boldsymbol{\Phi}^{-1}\left(\mathbf{u}\mathbb{E}[\mathbf{C}\boldsymbol{\alpha}]^\mathsf{T}\mathbb{E}[\mathbf{C}\boldsymbol{\alpha}] + \mathbb{E}[\mathbf{C}\,\mathcal{D}(\boldsymbol{\alpha})\mathbf{C}^\mathsf{T}]\right)\boldsymbol{\Phi}^{-1\mathsf{T}} \\ \mathbf{m} &= \boldsymbol{\Phi}^{-1}u\mathbb{E}[\mathbf{C}\boldsymbol{\alpha}] \end{aligned} \tag{17}$$

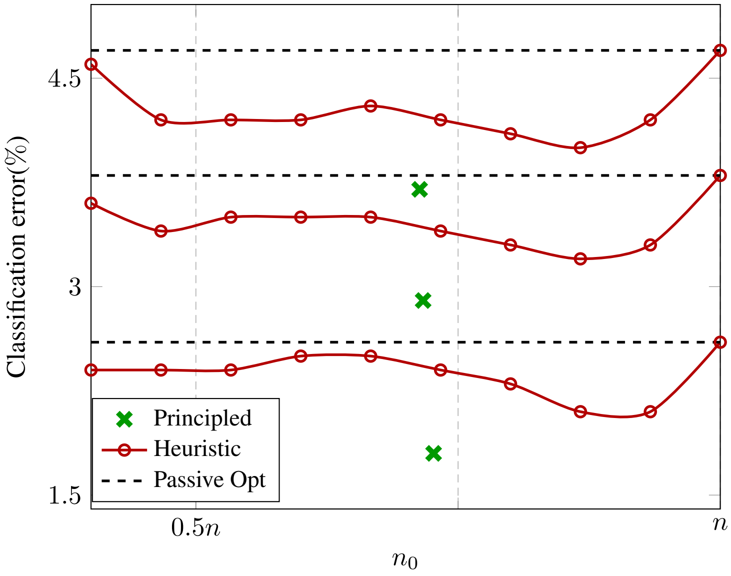

Figure 2: Empirical errors (averaged over 100 independent runs) of one-step active learning ($T = 1$) under a label budget of $n = 300$ on Gaussian mixture data of $p = 400$, $\mathbf{\Sigma} = \mathbf{I}_p - \mathbf{1}_{p,p}/2p$ and $\boldsymbol{\mu} = \sqrt{u}[\mathbf{1}_{p/2}; \mathbf{0}_{p/2}]/\sqrt{\frac{3}{4}p}$ for $u = \{4.5, 5, 6\}$ (from top to bottom), compared against the optimal passive learning performance (Theorem 1). Heuristic performance given by Algorithm 1 at various n_0, with $\ell_{qt} = (1-t)^2$ and $\mathcal{A} = (-0.05, 0.05)$. Principled learning achieved by Algorithm 2 with $\tilde{n}_0 = 1.5p$ and $\epsilon = 0.05$.

where $\boldsymbol{\alpha} = \{\alpha\}_{t=0}^T$, and

$$\begin{aligned} \mathbf{\Phi} &= -\,\mathbb{E}[\mathbf{C}\,\mathcal{D}(\boldsymbol{\alpha})(\mathbf{R} - \mathbf{m}\mathbf{1}_{T+1}^{\mathsf{T}})^{\mathsf{T}}]\mathbf{S}^{-1} \\ \mathbf{C} &= \mathbf{\Phi}(h_{\mathbf{\Phi}}(\mathbf{J}\otimes\mathbf{R}) - \mathbf{J}\otimes\mathbf{R}) \end{aligned}$$

with the operator $h_{\mathbf{\Phi}}(\cdot)$ defined as

$$\begin{aligned} h_{\mathbf{\Phi}}(\mathbf{M}) = \operatorname{argmin}_{\mathbf{M}'} \Bigg[&\sum_{q\le T'}\sum_{t\ge q} \ell_{tq}([\mathbf{M}]_{(t+1)(q+1)}) \\ &+ \frac{1}{2}\operatorname{tr}(\mathbf{M}'-\mathbf{M})^{\mathsf{T}}\mathbf{\Phi}(\mathbf{M}'-\mathbf{M})\Bigg] \end{aligned} \tag{18}$$

for any $\mathbf{M} \in \mathbb{R}^{(T+1)(T'+1)}$ with $T' \le T$, $\mathbf{J} \in \mathbb{R}^{(T+1)(T+1)}$ of $[\mathbf{J}]_{dd'} = \mathbb{1}_{d\ge d'}$, $\mathbf{R} = [\mathbf{r}_0, \ldots, \mathbf{r}_T]$ of $\mathbf{r}_0 \sim \times\mathcal{N}(\mathbf{m}, \mathbf{S})$ and

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

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. After obtaining $\mathbf{m}$ and $\mathbf{S}$, we can retrieve m_T and s_{TT} to predict the active learning performance as stated Theorem 2.

6.2 Technical Approach

The objective of this section is to give an overview of the technical approach. To focus on the key steps, we consider a symmetric model $\mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{\Sigma})$ with the bias term b set to zero and concentrate on investigating $\mathbf{w}$, and restrict our discussion to smooth loss functions.

Note first that the probability of correctly classifying a new observation $\mathbf{x}$ with the classifier $(\mathbf{w}_t, b_t)$ at the i-th iteration equals $Q(\tilde{m}_t/\sqrt{\tilde{s}_{tt}})$ with

$$\tilde{m}_t = \mathbf{w}_t^{\mathsf{T}}\boldsymbol{\mu}, \quad \tilde{s}_{tt} = \mathbf{w}_t^{\mathsf{T}}\mathbf{\Sigma}\mathbf{w}_t. \tag{19}$$

We define $\tilde{\mathbf{m}} = \{\tilde{m}_t\}_{t=0}^T$ and $\tilde{\mathbf{S}} = \{s_{tt'}\}_{t,t'=0}^{\mathsf{T}}$.

Our approach makes use of the leave-one-out procedure already applied in several analyses El Karoui et al. [2013], El Karoui [2013], Javanmard and Montanari [2015], Mai and Liao [2019] under settings other than active learning. It consists in introducing a leave-one-out version $\mathbf{w}_{(-i)}$ of $\mathbf{w}$, obtained by excluding the training sample $(\mathbf{x}_i, y_i)$ from the learning process, and rests on the premise that the difference $\mathbf{w} - \mathbf{w}_{(-i)}$ is small but highly correlated with the high-dimensional feature vector $\mathbf{x}_i$. This implies that $\mathbf{w}_{(-i)}^\mathsf{T}\mathbf{x}_i$ is not close to $\mathbf{w}^\mathsf{T}\mathbf{x}_i$ but has approximately the same distribution as $\mathbf{w}^\mathsf{T}\mathbf{x}$ for some new $\mathbf{x}$ (following the distribution (2) and independent of $\mathbf{w}$).

The study of active learning is more delicate due to the dependence with past iterations. Extra care should be taken when defining $\mathbf{w}_{(-i)T}$: simply removing $(\mathbf{x}_i, y_i)$ from the training set is not enough for obtaining $\mathbf{w}_{(-i)T}$ that is independent of $(\mathbf{x}_i, y_i)$, as other samples $(\mathbf{x}_j, y_j)$ that are queried after the $t_{[i]}$-th iteration where $(\mathbf{x}_i, y_i)$ is added to the training set are dependent of $(\mathbf{x}_i, y_i)$. We propose to construct $\mathbf{x}_{(-i)j}$ that are close to $\mathbf{x}_j$ and "roughly" independent of $\mathbf{x}_i$, in the sense that the leave-one-out classifiers $\mathbf{w}_{(-i)t}$ obtained on $(\mathbf{x}_{(-i)j}, y_j)$ can be treated as if they were independent of $\mathbf{x}_i$. Obviously, for $\mathbf{x}_j$ queried before or during the $t_{[i]}$-th iteration, it suffices to set $\mathbf{x}_j = \mathbf{x}_{(-i)j}$. To obtain $\mathbf{x}_{(-i)j}$ for other $\mathbf{x}_j$ queried after the $t_{[i]}$-th iteration, note that $\mathbf{x}_j$ (queried at the $t+1$-th iteration with $t \geq t_{[i]}$) can be written as

$$\mathbf{x}_j = y_j\boldsymbol{\mu} + \frac{1}{\tilde{s}_{tt}}(\sqrt{\tilde{s}_{tt}}\tilde{g}_j - \mathbf{w}_t^\mathsf{T}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{z}_j)\boldsymbol{\Sigma}\mathbf{w}_t + \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{z}_j$$

for $\tilde{g}_i$ with density $f(a) = f_{\mathcal{N}(0,1)}(a)\mathbb{1}_{\mathcal{A}}(\tilde{m}_t + \sqrt{\tilde{s}_{tt}}a)$ and $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$ independent of $\tilde{g}_i$. To remove the dependence of $\mathbf{x}_i$ brought by the correlation between $\mathbf{w}_t$ and $\mathbf{x}_i$, we define $\mathbf{x}_{(-i)j}$ as

$$\mathbf{x}_{(-i)j} = y_j\boldsymbol{\mu} + \frac{1}{\tilde{s}_{tt}}(\sqrt{\tilde{s}_{tt}}\tilde{g}_j - \mathbf{w}_t^\mathsf{T}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{z}_j)\boldsymbol{\Sigma}\mathbf{w}_{(-i)t} + \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{z}_j.$$

Cancelling the gradient of the empirical risk gives the following stationary point equation of $\mathbf{w}_T$

$$\sum_{q=0}^{T}\sum_{i=N_q-n_q+1}^{N_q} \ell'_{Tq}\left(y_i\mathbf{w}_T^\mathsf{T}\mathbf{x}_i\right)y_i\mathbf{x}_i = \mathbf{0}_p \tag{20}$$

A similar equation can also be established for $\mathbf{w}_{(-i)T}$. By taking the difference of the two equations, Taylor-expanding $\ell'_{qt}\left(y_j\mathbf{w}_T^\mathsf{T}\mathbf{x}_j\right)$ around $y_j\mathbf{w}_{(-i)T}^\mathsf{T}\mathbf{x}_{(-i)j}$ and multiplying $\mathbf{x}_i$ with the both sides of the equation (as conventionally done in the leave-one-out approach), we can get

$$\sum_{t=t_{[i]}}^{T}\phi_{Tt}y_i(\mathbf{w}_{(-i)t}^\mathsf{T}\mathbf{x}_i - \mathbf{w}_t^\mathsf{T}\mathbf{x}_i) \simeq \ell'_{Tt_{[i]}}\left(y_i(\mathbf{w}_T^\mathsf{T}\mathbf{x}_i)\right) \tag{21}$$

for ϕ_{Tt} some deterministic constants suggested by concentrations arguments. Define $\check{\mathbf{r}}_i = \{y_i\mathbf{w}_{(-i)t}^\mathsf{T}\mathbf{x}_i\}_{t=0}^{T}$, $\mathbf{j}_t \in \mathbb{R}^{T+1}$ with $[\mathbf{j}_t]_d = \mathbb{1}_{d>t}$. Establishing analogous equations to (21) for the past iterations $t = \{0, \dots, T-1\}$ allows us to obtain

$$\boldsymbol{\Phi}\left(h_{\boldsymbol{\Phi}}(\mathbf{j}_t \otimes \check{\mathbf{r}}_i) - \mathbf{j}_t \otimes \check{\mathbf{r}}_i\right) \simeq \begin{bmatrix} \mathbf{0}_{t_{[i]}} \\ -\ell'_{tt_{[i]}}(y_i\mathbf{w}_{(-i)t}^\mathsf{T}\mathbf{x}_i)\}_{t=t_{[i]}}^{T} \end{bmatrix}$$

where $h_{\boldsymbol{\Phi}}(\cdot)$ is as defined in (18) for some deterministic matrix $\boldsymbol{\Phi} = \{\phi_{tt'}\}_{t,t'=0}^{T}$. Plugging the above approximation into the stationary point equations (20) for the all iterations $t = \{0, \dots, T\}$ yields

$$[y_1\mathbf{x}_1, \dots, y_n\mathbf{x}_n]\check{\mathbf{C}}^\mathsf{T} \simeq \mathbf{0}_{p,(T+1)} \tag{22}$$

with $\check{\mathbf{C}} = [\check{\mathbf{c}}_1, \dots, \check{\mathbf{c}}_n] \in \mathbb{R}^{(\mathsf{T}+1)\times n}$ of

$$\check{\mathbf{c}}_i = \boldsymbol{\Phi}\left(h_{\boldsymbol{\Phi}}(\mathbf{j}_t \otimes \check{\mathbf{r}}_i) - \mathbf{j}_t \otimes \check{\mathbf{r}}_i\right).$$

We define

$$\boldsymbol{\omega}_i = \left(\mathbf{I}_p - \mathbf{P}_{(-i)}\right)\boldsymbol{\Sigma}^{-\frac{1}{2}}(y_i\mathbf{x}_i - \boldsymbol{\mu})$$

with $\mathbf{P}_{(-i)}$ the projection matrix onto the space spanned by $\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{w}_{(-i)t}$, $t \in \{0, \dots, T\}$. It is easy to check that $\mathrm{Cov}\left[\boldsymbol{\omega}_i, \mathbf{x}_i^\mathsf{T}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{w}_{(-i)t}\right] = \mathbf{0}_{T+1}$ for all t. We remark by Gaussian invariance that $\boldsymbol{\omega}_i$ is independent of all $\mathbf{x}_i^\mathsf{T}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{w}_{(-i)t}$ and therefore also independent of $\check{\mathbf{c}}_i$. As the leave-one-out argument $\mathbf{w}_{(-i)t} \simeq \mathbf{w}_t$ implies

$$\mathbf{P}_{(-i)} \simeq \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{W}\tilde{\mathbf{S}}^{-1}\mathbf{W}_{(-i)}^\mathsf{T}\boldsymbol{\Sigma}^{\frac{1}{2}}$$

where $\mathbf{W} = [\mathbf{w}_0, \ldots, \mathbf{w}_T]$ and $\mathbf{W}_{(-i)} = [\mathbf{w}_{(-i)0}, \ldots, \mathbf{w}_{(-i)T}]$, we can rewrite (22) as

$$\frac{1}{p}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}\mathbf{1}_n^\mathsf{T} + \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Omega}\right)\check{\mathbf{C}}^\mathsf{T} \simeq \mathbf{W}\tilde{\boldsymbol{\Phi}}^\mathsf{T} \tag{23}$$

where $\boldsymbol{\Omega} = [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}]$ and $\tilde{\boldsymbol{\Phi}}^\mathsf{T} = -\frac{1}{p}\mathbf{S}^{-1}(\check{\mathbf{R}} - \tilde{\mathbf{m}}\mathbf{1}_n^\mathsf{T})\check{\mathbf{C}}^\mathsf{T}$ with $\check{\mathbf{R}} = [\check{\mathbf{r}}_1, \ldots, \check{\mathbf{r}}_n]$. The above decomposition is interesting for the reason that $\boldsymbol{\Omega}$ can be treated conveniently as if it was independent of $\check{\mathbf{C}}$. It can be deduced from (23) and the intuition that $\mathbf{w}_t - \mathbf{w}_{(-i)t}$ is highly correlated with $\mathbf{x}_i$ that

$$\tilde{\boldsymbol{\Phi}}(\mathbf{W} - \mathbf{W}_{(-i)})^\mathsf{T} y_i \mathbf{x}_i \simeq \frac{1}{p}\boldsymbol{\omega}_i^\mathsf{T}\boldsymbol{\Sigma}^{-\frac{1}{2}} y_i \mathbf{x}_i \check{\mathbf{c}}_i \simeq \check{\mathbf{c}}_i,$$

which entails $\tilde{\boldsymbol{\Phi}} \simeq \boldsymbol{\Phi}$ according to (21).

In the regime of abundant data summations over $i \in \{1, \ldots, n\}$ tend to converge to their expectation. We have then from (23) that

$$\begin{aligned}
\tilde{\boldsymbol{\Phi}}\tilde{\mathbf{m}} \simeq& \frac{u}{p}\mathbb{E}[\check{\mathbf{C}}\mathbf{1}_n] = u\mathbb{E}[\tilde{\mathbf{C}}\alpha] \\
\tilde{\boldsymbol{\Phi}}\tilde{\mathbf{S}}\tilde{\boldsymbol{\Phi}}^\mathsf{T} \simeq& \frac{u}{p^2}\mathbf{u}\mathbb{E}[\check{\mathbf{C}}\mathbf{1}_n]^\mathsf{T}\mathbb{E}[\check{\mathbf{C}}\mathbf{1}_n] + \frac{1}{p}\mathbb{E}[\check{\mathbf{C}}\check{\mathbf{C}}^\mathsf{T}] \\
=& \mathbf{u}\mathbb{E}[\tilde{\mathbf{C}}\boldsymbol{\alpha}]^\mathsf{T}\mathbb{E}[\tilde{\mathbf{C}}\boldsymbol{\alpha}] + \mathbb{E}[\tilde{\mathbf{C}}\,\mathcal{D}(\boldsymbol{\alpha})\tilde{\mathbf{C}}^\mathsf{T}]
\end{aligned} \tag{24}$$

for $\boldsymbol{\alpha} = \{\alpha_t\}_{t=0}^T$, $\tilde{\mathbf{C}}$ random matrix with its $t+1$-th column vector having the same distribution as any $\check{\mathbf{c}}_i$ such that $\mathbf{x}_i$ is queried at the t-th iteration. The fist equation of (24) is obtained by multiplying (23) with $\boldsymbol{\mu}$ and taking the expectation at the right side, and the second by multiplying (23) with its transpose and also taking the expectation at the right side. The system (17) of equations is thus retrieved by letting $(\mathbf{m}, \mathbf{S}, \boldsymbol{\Phi})$ be the deterministic limit of $(\tilde{\mathbf{m}}, \tilde{\mathbf{S}}, \tilde{\boldsymbol{\Phi}})$ and $(\tilde{\mathbf{C}}, \tilde{\mathbf{R}})$ the asymptotic equivalent of $(\mathbf{C}, \mathbf{R})$.

7 Conclusion

In this article we provided the first precise performance analysis for confidence-based active learning that (i) fascilitates understanding the joint effect of data distribution and hyperparameters, (ii) gives a performance prediction that is exact for sufficiently large dimensional data. Based on our analysis, we were able to demonstrate provable performance gains for confidence-based sampling over passive learning. We also derived new insights regarding the optimal query strategy, the effect of initial classifier and the reliability of large-batch query. Finally, we showed that our framework provides practical guidelines for addressing the challenging problem of hyperparameter tuning in active learning.

Our framework can be potentially extended in several directions. First, our performance guarantees for confidence-based active learning may be sharpened by deriving performance bounds that are dependent on the query iteration number (the current ones hold universally for any iteration number greater than 1). Second, following up on our discovery that labelling the most uncertain data is not always the best choice, better practical designs can be envisioned with an optimized query strategy, Finally, with some additional effort, our analysis can be adapted to a wider range of scenarios in terms of data distribution and algorithmic framework, by building upon the flexible technical approach proposed in Section 6.2. As the first high dimensional asymptotic analysis of active learning, our study may have broader implications for the study of algorithms that have a similar iterative nature as confidence-based sampling.

References

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer, 1994.

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*, 9(2):209–223, 2017.

Richard Segal, Ted Markowitz, and William Arnold. Fast uncertainty sampling for labeling large e-mail corpora. In *CEAS*. Citeseer, 2006.

Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *ICML*, volume 2, page 6. Citeseer, 2000.

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*, pages 953–960, 2009.

Alexander Liu, Goo Jun, and Joydeep Ghosh. A self-training approach to cost sensitive uncertainty sampling. *Machine learning*, 76(2-3):257–270, 2009.

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

Yazhou Yang and Marco Loog. A benchmark and comparison of active learning for logistic regression. *Pattern Recognition*, 83:401–415, 2018.

Maria E Ramirez-Loaiza, Manali Sharma, Geet Kumar, and Mustafa Bilgic. Active learning: an empirical study of common baselines. *Data mining and knowledge discovery*, 31(2):287–313, 2017.

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*, pages 1070–1079, 2008.

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

Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. *Advances in neural information processing systems*, 18:235–242, 2005.

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

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems*, pages 442–450, 2014.

Tzu-Kuo Huang, Alekh Agarwal, Daniel J Hsu, John Langford, and Robert E Schapire. Efficient and parsimonious agnostic active learning. In *Advances in Neural Information Processing Systems*, pages 2755–2763, 2015.

Alina Beygelzimer, Daniel J Hsu, John Langford, and Tong Zhang. Agnostic active learning without constraints. In *Advances in neural information processing systems*, pages 199–207, 2010.

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 49–56, 2009.

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*, page 201307842, 2013.

Romain Couillet, Florent Benaych-Georges, et al. Kernel spectral clustering of large dimensional data. *Electronic Journal of Statistics*, 10(1):1393–1454, 2016.

Hanwen Huang. Asymptotic behavior of support vector machine for spiked population model. *Journal of Machine Learning Research*, 18(45):1–21, 2017.

Edgar Dobriban, Stefan Wager, et al. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

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

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

Zhenyu Liao and Romain Couillet. A large dimensional analysis of least squares support vector machines. *IEEE Transactions on Signal Processing*, 67(4):1065–1074, 2019.

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

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*, 68:2464–2479, 2020.

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

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

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

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*, 19(1):3074–3100, 2018.

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

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

Noureddine El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

Adel Javanmard and Andrea Montanari. De-biasing the Lasso: Optimal Sample Size for Gaussian Designs. *arXiv e-prints*, art. arXiv:1508.02757, August 2015.