Asymptotic Analysis of Objectives based on Fisher Information in Active Learning

Jamshid Sourati∗         Murat Akcakaya†       Todd K. Leen‡       Deniz Erdogmus§
Jennifer G. Dy¶

May 31, 2016

Abstract

Obtaining labels can be costly and time-consuming. Active learning allows a learning algorithm to intelligently query samples to be labeled for efficient learning. Fisher information ratio (FIR) has been used as an objective for selecting queries in active learning. However, little is known about the theory behind the use of FIR for active learning. There is a gap between the underlying theory and the motivation of its usage in practice. In this paper, we attempt to fill this gap and provide a rigorous framework for analyzing the existing FIR-based active learning methods. In particular, we show that FIR can be asymptotically viewed as an upper bound of the expected variance of the log-likelihood ratio. Additionally, our analysis suggests a unifying framework that not only enables us to make theoretical comparisons among the existing querying methods based on FIR, but also gives more insight into the development of new active learning approaches based on this objective.

1 Introduction

In supervised learning, a learner is a model-algorithm pair that is optimized to (semi) automatically perform tasks, such as classification, or regression using information provided by an external source (oracle). In passive learning, the learner has no control over the information given. In active learning, the learner is permitted to query certain types of information from the oracle. Usually there is a cost associated with obtaining information from an oracle; therefore an active learner will need to maximize the information gained from queries within a fixed budget or minimize the cost of gaining a desired level of information. A majority of the existing algorithms restrict to the former problem, to get the most efficient trained learner by querying a fixed amount of knowledge [19, 40].

∗Department of Electrical and Computer Engineering, Northeastern University, Boston MA. E-mail: sourati@ece.neu.edu
†Department of Electrical and Computer Engineering, University of Pittsburgh. E-mail: akcakaya@pitt.edu
‡Georgetown University. E-mail: todd.leen@georgetown.edu
§Department of Electrical and Computer Engineering, Northeastern University, Boston MA. E-mail: erdogmus@ece.neu.edu
¶Department of Electrical and Computer Engineering, Northeastern University, Boston MA. E-mail: jdy@ece.neu.edu
Active learning is the process of coupled querying/learning strategies. In such an algorithm, one needs to specify a query quality measure in terms of the learning method that uses the new information gained at each step of querying. For instance, information theoretic measures are commonly employed in classification problems to choose training samples whose class labels, considered as random variables, are most informative with respect to the labels of the remaining unlabeled samples. This family of measures is particularly helpful when probabilistic approaches are used for classification. Among these objectives, Fisher information criterion is very popular due to its relative ease of computation compared to other information theoretic objectives, desirable statistical properties and existence of effective optimization techniques. However, as we discuss in this manuscript, this objective is not well-studied in the classification context and there seems to be a gap between the underlying theory and the motivation of its usage in practice.

This paper is an attempt to fill this gap and also provide a rigorous framework for analyzing the existing querying methods based on Fisher information.

In order to better categorize the existing active learning techniques, we characterize the process of constructing a classifier in three steps as follows: (1) choosing the loss and risk functions, (2) building a decision rule that minimizes the risk, and (3) modeling the discriminant functions of the decision rule. For instance, choosing the simple 0/1 loss and its a posteriori expectation as the risk, incurs the Bayes rule as the optimal decision [13], where the discriminant function is the posterior distribution of the class labels given the covariates. For this type of risk, discriminative models that directly parametrize the posteriors, such as logistic regression, are popularly used to learn the discriminant functions [5]. Each active learning algorithm can be broken into the following sub-problems:

(i) **(Query Selection)** Sampling a set of covariates \(\{x_1, ..., x_n\}\) from the training marginal \(\mathcal{M}_1\), whose labels \(\{y_1, ..., y_n\}\) are to be requested from an external source of knowledge (the *oracle*). The queried covariates together with their labels form the training data set.

(ii) **(Inference)** Estimating parameters of the posterior model based on the training data set formed in the previous step.

(iii) **(Prediction)** Making decisions regarding class labels of the test covariates sampled from the test marginal.

These three steps can be carried out iteratively. Note that the query selection sub-problem is formulated in terms of the distribution from which the queries will be drawn. Ideally, queries (or the query distribution) are chosen such that they increase the expected quality of the classification performance measured by a particular objective function. This objective can be constructed from two different perspectives: based on the accuracy of the parameter inference or the accuracy of label prediction. In the rest of the manuscript, accordingly, we refer to the algorithms that use these two types of objectives as *inference-based* or *prediction-based* algorithms, respectively.

Most of the inference-based querying algorithms in classification aim to choose queries that maximize the expected change in the objective of the inference step [24, 42]. Although there are more recent algorithms for bound minimizing of the parameter loss function [7, 21]. On the other hand, the wide range of studies in prediction-based active learning includes a more varied

---

1 Throughout this paper, marginal distribution or simply distribution refers to the distribution of covariates, while joint distribution is used for pairs of the covariates and their class labels.
set of objectives: for instance, the prediction error probability, variance of the predictions, uncertainty of the learner with respect to the unknown labels as evaluated by the entropy function, mutual information, margin of the samples with respect to the trained hyperplanar discriminant function or Fisher information criterion.

In this manuscript, we focus on the Fisher information criterion used in classification active learning algorithms. These algorithms use a scalar function of the Fisher information matrices computed for parametric models of training and test marginals. In the classification context, this scalar is sometimes called Fisher information ratio (FIR) and its usage is motivated by older attempts in optimal experiment design for statistical regression methods.

Among the existing FIR-based classification querying methods, only the very first one proposed by Zhang and Oles approached the FIR objective from a parameter inference point of view. Using a maximum likelihood estimator (MLE), they claimed (with the proof skipped) that FIR is asymptotically equal to the expectation of the log-likelihood ratio with respect to both test and training samples (see sub-problem). Later on, Hoi et al. and Hoi et al., inspired by Zhang and Oles, used FIR in connection with a logistic regression classifier with the motivation of decreasing the labels’ uncertainty and hence the prediction error. Settles and Craven employed this objective with the same motivation, but using a different approximation and optimization technique. More recently, Chaudhuri et al. showed that the Fisher trace objective is the upper-bound of the log-likelihood error and therefore can be used as a bound-minimizing objective in query selection. However, this result is restricted to pool-based scenarios where the test distribution is assumed to be a uniform probability mass function (PMF) over the pool, and also the restrictive assumption that the Hessian matrix of the model is independent of the class labels.

Here, similar to Zhang and Oles but with a more expanded and different derivation, we discuss a novel theoretical result based on which FIR is related to an MLE-based inference step for a large number of training data. More specifically, under certain regularity conditions required for consistency of MLE and in the absence of model mis-specification, we show that FIR can be viewed as an upper bound for the expected variance of the asymptotic distribution of the log-likelihood ratio. Unlike Chaudhuri et al., our derivations are based on the generic regularity and identifiability properties for the distributions and made no restrictive assumptions over the test marginal and the format of FI matrices.

There are two practical issues in employing FIR as a query selection objective: its computation and optimization. First, computing the Fisher information matrices is usually intractable, except for very simple distributions; also FIR depends on the true marginal, which is usually unknown. Therefore, even if the computations are tractable, approximations have to be used for evaluating FIR. Second, the optimization of FIR is straightforward only if a single query is to be selected per iteration, or when the optimization has continuous domain (e.g., optimizing to get the real parameters of the query marginal). However, the optimization becomes NP-hard when multiple queries are to be selected from a countable set of unlabeled samples (pool-based batch active learning). Heuristics have been used to approximate such combinatorial optimization, such as greedy methods and relaxation to continuous domains. Another strategy is to take advantage of monotonic submodularity of the objective set functions. If the objective is

\footnote{Prediction error probability is indeed the frequentist risk function of 0/1 loss, and is also known as generalization error.}
shown to be monotonically submodular, efficient greedy algorithms can be used for optimization with guaranteed tight bounds \[2, 8, 30\]. Regarding FIR, Hoi et al. \[27\] proved that, when a logistic regression model is used, a Monte-Carlo simulation of this objective is a monotone and submodular set function in terms of the queries.

In addition to our theoretical contribution in asymptotically relating FIR to the log-likelihood ratio, we clarify the differences between some of the existing FIR-based querying methods according to the techniques that they use to address the evaluation and optimization issues. Furthermore, we show that monotonicity and submodularity of Monte-Carlo approximation of FIR can be extended from logistic regression models to \textit{any} discriminative classifier. Here is a summary of our contributions in this paper:

- Establishing a relationship between the Fisher information matrix of the query distribution and the asymptotic distribution of the log-likelihood ratio (section 4.1);

- Showing that FIR can be viewed as an upper bound of the expected asymptotic variance of the log-likelihood ratio without restrictive assumptions over the test/training distributions. This implies that minimizing FIR, as an active learning objective, is asymptotically equivalent to upper-bound minimization of the expected variance of the log-likelihood ratio, as a measure of inference performance (section 4.1);

- Discussing different methods for coping with practical issues in evaluating and optimizing FIR (section 4.3), and accordingly providing a unifying framework for existing FIR-based active learning methods (section 4.4).

- Proving submodularity for the Monte-Carlo simulation of FIR under \textit{any} discriminative classifier, assuming a pool-based active learning, and therefore having access to approximations of Fisher information matrices of both test and training distributions (section 4.4).

Before going through the main discussion in section 4, we formalize our classification model assumptions, set the notations and review the basics and some of the key properties of our inference method, maximum likelihood estimation, in sections 2 and 3. The statistical background required to follow the remaining sections is given in Appendix A.

2 The Framework and Assumptions

In this paper, we deal with classification problems, where each covariate, represented by a feature vector \( \mathbf{x} \) in vector space \( X \), is associated with a numerical class label \( y \). Assuming that there are \( 1 < c < \infty \) classes, \( y \) can take any integer among the set \( \{1, \ldots, c\} \). Suppose that the pairs \((\mathbf{x}, y)\) are distributed according to a parametric joint distribution \( p(\mathbf{x}, y|\theta) \), with the parameter space denoted by \( \Omega \subseteq \mathbb{R}^d \). Using a set of observed pairs as the training data, \( \mathcal{L}_n := \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \), we can estimate \( \theta \) and predict the class labels of the unseen test samples, e.g. by maximizing \( p(y|\mathbf{x}, \theta) \). In active learning, the algorithm is permitted to take part in designing \( \mathcal{L}_n \) by choosing a set of data points \( \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \), for which the class labels are then generated using an external oracle.

In addition to the framework described in the last section (see subproblems (i) to (iii)), we make the following assumptions regarding the oracle, our classification model and the underlying data distribution:
(A0). The dependence of the joint distribution to the parameter $\theta$ comes only from the class-
conditional distribution and the marginal distribution does not depend on $\theta$, that is:
\[
p(x, y | \theta) = p(y | x, \theta)p(x).
\]

Zhang and Oles \[48\] referred to joint distributions with such parameter dependence as
type-II models, as opposed to type-I models which have parameter dependence in both
class conditionals and marginal. They argue that active learning is more suitable for
type-II models. Moreover, maximizing the joint with respect to the parameter vector in
this model, becomes equivalent to maximizing the posterior $p(y | x, \theta)$ (inference step in
sub-problem (ii)).

(A1). (Identifiability) : The joint distribution $P_\theta$ (whose density is given by $p(x, y | \theta)$) is identi-
ifiable for different parameters. Meaning that for every distinct parameter vectors $\theta_1$ and
$\theta_2$ in $\Omega$, $P_{\theta_1}$ and $P_{\theta_2}$ are also distinct. That is
\[
\forall \theta_1 \neq \theta_2 \in \Omega \exists A \subseteq X \times \{1, ..., c\} \ s.t. \ P_{\theta_1}(A) \neq P_{\theta_2}(A).
\]

(A2). The joint distribution $P_\theta$ has common support for all $\theta \in \Omega$.

(A3). (Model Faithfulness) : For any $x \in X$, we have access to an oracle that generates a label
$y$ according to the conditional $p(y | x, \theta_0)$. That is, the posterior parametric model matches
the oracle distribution. We call $\theta_0$ the true model parameter.

(A4). (Training joint) : The set of observations in $L_n := \{(x_1, y_1), ..., (x_n, y_n)\}$ are drawn inde-
dependently from the training/proposal/query joint distribution of the form $p(y | x, \theta_0)q(x)$
where $q$ is the training marginal with no dependence on the parameter.

(A5). (Test joint) : The unseen test pairs are distributed according to the test/true joint
distribution of the form $p(y | x, \theta_0)p(x)$ where $p$ is the test marginal with no dependence
on the parameter.

(A6). (Differentiability) : The log-conditional log $p(y | x, \theta)$ is of class $C^3(\Omega)$ as a function of $\theta$
and for all $(x, y) \in X \times \{1, ..., c\}$.

(A7). The parameter space $\Omega$ is compact and there exists an open ball around the true parameter
of the model $\theta_0 \in \Omega$.

(A8). (Invertibility) : The Fisher information matrix (reviewed in section 3.2) of the joint
distribution is positive definite and therefore invertible for all $\theta \in \Omega$, and for any type of
marginal that is used under assumption (A0).

Regarding assumptions (A4) and (A5), note that the training and test marginals are not neces-
sarily equal. The test marginal is usually not known beforehand and $q$ cannot be set equal
to $p$ in practice, hence $q$ can be viewed as a proposal distribution. Such inconsistency is what
Shimodaira \[43\] called covariate shift in distribution. In the remaining sections of the report, we
use subscripts $p$ and $q$ for the statistical operators that consider $p(x)$ and $q(x)$ as the marginal
in the joint distribution, respectively.

\[\text{We say that a function } f : X \rightarrow Y \text{ is of } C^p(X), \text{ for an integer } p > 0, \text{ if its derivatives up to } p\text{-th order exist and are continuous at all points of } X.\]
3 Background

Here, we provide a short review of maximum likelihood estimation (MLE) as our inference method, and briefly introduce Fisher information of a parametric distribution. These two basic concepts enable us to explain some of the key properties of MLE, upon which our further analysis of FIR objective relies. Note that our focus in this section is on sub-problem (ii) with the assumptions listed above.

3.1 Maximum Likelihood Estimation

In this section, we review maximum likelihood estimation in the context of classification problem. Given a training data set \( L_n = \{(x_1, y_1), ..., (x_n, y_n)\} \), a maximum likelihood estimate (MLE) is obtained by maximizing the log-likelihood function over all pairs inside \( L_n \), with respect to the parameter \( \theta \):

\[
\hat{\theta}_n = \arg \max_{\theta} \log p(L_n | \theta), \tag{2}
\]

Under the assumptions \([A0]\) and \([A4]\) the optimization in (2) can be written as

\[
\hat{\theta}_n = \arg \max_{\theta} \sum_{i=1}^{n} \log p(y_i | x_i, \theta), \tag{3}
\]

Equation (3) shows that MLE does not depend on the marginal when using type-II model. Hence, in our analysis we focus on the conditional log-likelihood as the classification objective, and simply call it the log-likelihood function when viewed as a function of the parameter vector \( \theta \), for any given pair \((x, y) \in X \times \{1, ..., c\}\):

\[
\ell(\theta; x, y) := \log p(y | x, \theta). \tag{4}
\]

Moreover, for any set of pairs independently generated from the joint distribution of the training data, such as \( L_n \) mentioned in \([A4]\) the log-likelihood function will be:

\[
\ell(\theta; L_n) = \sum_{i=1}^{n} \ell(\theta; x_i, y_i) = \sum_{i=1}^{n} \log p(y_i | x_i, \theta). \tag{5}
\]

hence the MLE can be rewritten as

\[
\hat{\theta}_n = \arg \max_{\theta} \sum_{i=1}^{n} \ell(\theta; x_i, y_i). \tag{6}
\]

Doing this maximization usually involves the computation of the stationary points of the log-likelihood, which requires calculating \( \nabla_{\theta} \ell(\theta; L_n) = \sum_{i=1}^{n} \nabla_{\theta} \ell(\theta; x_i, y_i) \). For models assumed in \([A0]\) each of the derivations in the summation is equal to the score function defined as the gradient of the joint log-likelihood:

\[
\nabla_{\theta} \ell(\theta; x, y) = \nabla_{\theta} \log p(y | x, \theta) = \nabla_{\theta} \log p(x, y | \theta), \tag{7}
\]
Equation (7) implies that the score will be the same no matter whether we choose the training or test distribution as our marginal. Furthermore, under regularity conditions (A6), the score is always a zero-mean random variable.

Finally, using MLE to estimate $\hat{\theta}_n$, class label of a test sample $x$ will be predicted as the class with the highest log-likelihood value:

$$\hat{y}(x) = \arg \max_y \ell(\hat{\theta}_n; x, y).$$

(8)

3.2 Fisher Information

Here we give a very short introduction to Fisher information. More detailed descriptions about this well-known criterion can be found in various textbooks, such as Lehmann and Casella [31].

Fisher information of a parametric distribution is a measure of information that the samples generated from that distribution provide regarding the parameter. It owes part of its importance to the Cramér-Rao Theorem (see Appendix A.2, Theorem 12), which guarantees a lower-bound for the covariance of the parameter estimators.

Fisher information, denoted by $I(\theta)$, is defined as the expected value of the outer-product of the score function with itself, evaluated at some $\theta \in \Omega$. In our classification context, taking the expectation with respect to the training or test distributions gives us the training or test Fisher information criteria, respectively:

$$I_q(\theta) := E_q \left[ \nabla_{\theta} \ell(\theta; x, y) \cdot \nabla_{\theta}^\top \ell(\theta; x, y) \right]$$

$$I_p(\theta) := E_p \left[ \nabla_{\theta} \ell(\theta; x, y) \cdot \nabla_{\theta}^\top \ell(\theta; x, y) \right]$$

(9)

Here, we focus on $I_q$ to further explain Fisher information criterion. Our descriptions here can be directly generalized to $I_p$ as well. First, note that from equation (7) and that the score function is always zero-mean, one can reformulate the definition as:

$$I_q(\theta) = E_q \left[ \nabla_{\theta} \ell(\theta; x, y) \cdot \nabla_{\theta}^\top \ell(\theta; x, y) \right] = \text{Cov}_q \left[ \nabla_{\theta} \ell(\theta; x, y) \right]$$

(10)

Under the differentiability conditions (A6), it is easy to show that we can also write the Fisher information in terms of the Hessian matrix of the log-likelihood:

$$I_q(\theta) = -E_q \left[ \nabla_{\theta}^2 \ell(\theta; x, y) \right]$$

(11)

Recall that the subscript $q$ in equations (10) and (11) indicates that the expectations are taken with respect to the joint distribution that uses $q(x)$ as the marginal, that is $p(x, y|\theta) = q(x)p(y|x, \theta)$. Expansion of the expectation in (11) results

$$I_q(\theta) = -E_q(x) \left[ \sum_y p(y|x, \theta) \cdot \nabla_{\theta}^2 \ell(\theta; x, y) \right] d x$$

(12)

4Score function is actually zero-mean even under weaker regularity conditions.
3.3 Some Properties of MLE

In this section, we formalize some of the key properties of MLE, which make this estimator popular in various fields. They are also very useful in the theoretical analysis of FIR, provided in section 4.1. More detailed descriptions of these properties, together with the proofs that are skipped here, can be found in different sources, such as Wasserman [45] and Lehmann and Casella [31].

Note that a full understanding of the properties described in this section requires the knowledge of different modes of statistical convergence, specifically, convergence in probability (\( \xrightarrow{P} \)), and convergence in law (\( \xrightarrow{L} \)). A brief overview of these concepts are given in Appendix A.

**Theorem 1** (Lehmann and Casella [31], Theorem 5.1). If the assumptions (A0) to (A7) hold, then there exists a sequence of solutions \( \{\hat{\theta}_n^*\}_{n=1}^{\infty} \) to \( \nabla_\theta \ell(\theta; L_n) = 0 \) that converges to the true parameter \( \theta_0 \) in probability.

Note that Theorem 1 does not imply that convergence holds for any sequence of MLEs. Hence, if there are multiple solutions to equation \( \nabla_\theta \ell(\theta; L_n) = 0 \) (the equation to solve for finding the stationary points) for every \( n \), it is not obvious which root to select as \( \hat{\theta}_n^* \) to sustain the convergence. Therefore, while consistency of the MLE is guaranteed for models with a unique root of the score function evaluated at \( L_n \), it is not trivial how to build a consistent sequence when multiple roots exist. Here, in order to remove this ambiguity, we assume that either the roots are unique, become asymptotically unique, or we have access to an external procedure guiding us to select the proper roots so that \( \hat{\theta}_n \xrightarrow{P} \theta_0 \). We will denote the selected roots the same as \( \hat{\theta}_n \) from now on.

**Theorem 2** (Lehmann and Casella [31], Theorem 5.1). Let \( \hat{\theta}_n \) be the maximum likelihood estimator based on the training data set \( L_n \). If the assumptions (A0) to (A8) hold, then the MLE \( \hat{\theta}_n \) has a zero-mean normal asymptotic distribution with the covariance equal to the inverse Fisher information matrix, and with the convergence rate of \( 1/2 \):

\[
\sqrt{n} \cdot (\hat{\theta}_n - \theta_0) \xrightarrow{L} \mathcal{N}(0, I_q(\theta_0)^{-1}) \tag{13}
\]

Theorems 2 and Cramér-Rao bound (see Appendix A), together with the consistency assumption, i.e. \( \hat{\theta}_n \xrightarrow{P} \theta_0 \), imply that MLE is an asymptotically efficient estimator with the efficiency equal to the training Fisher information. One can rewrite (13) as

\[
\sqrt{n} \cdot I_q(\theta_0)^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{L} \mathcal{N}(0, I_d) \tag{14}
\]

In the following corollary, we see that if we substitute \( I_q(\theta_0) \) with \( I_q(\hat{\theta}_n) \), the new sequence still converges to a normal distribution:

**Corollary 1** (Wasserman [45], Theorem 9.18). Under the assumptions of Theorem 2, we get

\[
\sqrt{n} \cdot (\hat{\theta}_n - \theta_0) \xrightarrow{L} \mathcal{N}\left(0, I_q(\hat{\theta}_n)^{-1}\right) \tag{15}
\]
Note that equation (15) simply means that $\mathcal{N}(0, \mathbf{I}_q(\hat{\theta}_n)^{-1})$ is a close approximation of the distribution of the random variable $\sqrt{n} \cdot (\hat{\theta}_n - \theta_0)$ for large values of $n$. In similar situations appeared in the rest of this paper, whenever there exists a density with parameters depending on $\hat{\theta}_n$ on the right-hand-side of $\frac{L}{\sqrt{n}}$, we mean that the distribution of the random variable in the left-hand-side can be approximated by that, with an increasing accuracy as $n$ gets larger. This abuse of notation significantly simplifies the representation of our results (mostly in Corollary 2).

4 Fisher Information Ratio as an Objective of Active Learning for Classification

In this section, first we give a theoretical analysis to relate FIR to the asymptotic distribution of the parameter log-likelihood ratio (section 4.1). In section 4.2, we show that the derived theoretical results can be used in an active learning application. Then, we discuss practical issues in evaluating and optimizing the querying objective (section 4.3), and accordingly categorize the previous FIR-based active learning methods (section 4.4).

4.1 Theoretical Analysis

Recall that the estimated parameter $\hat{\theta}_n$ is obtained from a given proposal distribution $q(x)$. The log-likelihood ratio function, at a given pair $(x, y)$, is defined as:

$$\ell(\hat{\theta}_n; x, y) - \ell(\theta_0; x, y).$$

(16)

This ratio can be viewed as an example of the classification loss function whose expectation with respect to the test joint distribution of $x$ and $y$, results in the discrepancy between the true conditional $p(y | x, \theta_0)$ and MLE conditional $p(y | x, \hat{\theta}_n)$ [33]. Here, we analyze this measure asymptotically as $(n \to \infty)$. Primarily, note that based on continuity of the log-likelihood function (A6) and consistency of MLE (Theorem 1), equation (16) converges in probability to zero for any $(x, y)$.

Furthermore, equation (16) is dependent on both the true marginal $p(x)$ (through the test pairs, where it should be evaluated) and the proposal marginal $q(x)$ (through the MLE $\hat{\theta}_n$). In the classification context, Zhang and Oles [48] claimed that the expected value of this ratio with respect to both marginals converges to $\text{tr} \left[ \mathbf{I}_q(\theta_0)^{-1} \mathbf{I}_p(\theta_0) \right]$ with the convergence rate equal to unity. In the scalar case, $\text{tr} \left[ \mathbf{I}_q(\theta_0)^{-1} \mathbf{I}_p(\theta_0) \right]$ is equal to the ratio of the Fisher information of the true and proposal distributions, the reason why it is sometimes referred to as the Fisher information ratio [41]. This objective have been widely studied in linear and non-linear regression problems [10, 16, 20, 32, 33]. However, it is not as fully analyzed in classification.

Zhang and Oles [48] and many papers following them [27, 28, 41], used this function as an asymptotic objective in active learning to be optimized with respect to the proposal $q$. Here, we show that this objective can also be viewed as an upper bound for the expected variance of the asymptotic distribution of (16).

First, we investigate the asymptotic distribution of the log-likelihood ratio in two different cases:

**Theorem 3.** If the assumptions (A0) to (A8) hold, then, at any given $(x, y) \in X \times \{1, \ldots, c\}$:
(I) In case $\nabla_\theta \ell(\theta_0; x, y) \neq 0$, the log-likelihood ratio follows an asymptotic normality with convergence rate equal to $1/2$. More specifically

$$\sqrt{n} \left( \ell(\hat{\theta}_n; x, y) - \ell(\theta_0; x, y) \right) \xrightarrow{L} \mathcal{N} \left( 0, \text{tr} \left[ \nabla_\theta \ell(\theta_0; x, y) \cdot \nabla_\theta \ell(\theta_0; x, y) \cdot I_q(\theta_0)^{-1} \right] \right).$$ (17)

(II) In case $\nabla_\theta \ell(\theta_0; x, y) = 0$ and $\nabla_\theta^2 \ell(\theta_0; x, y)$ is non-singular, the asymptotic distribution of the log-likelihood ratio is a mixture of first-order Chi-square distributions, and the convergence rate is one. More specifically:

$$n \cdot \left( \ell(\hat{\theta}_n; x, y) - \ell(\theta_0; x, y) \right) \xrightarrow{L} \sum_{i,j=1}^{d} \left[ \gamma_{ij} \cdot \chi^2_1 \right]$$ (18)

where

$$\gamma_{ij} = \left[ I_q(\theta_0)^{-1/2} \nabla_x^2 \ell(\theta_0; x, y) I_q(\theta_0)^{-1/2} \right]_{ij}$$ (19)

Proof. Due to assumptions (A0–A7), Theorem 2 holds and therefore we have $\sqrt{n} \cdot (\hat{\theta}_n - \theta_0) \xrightarrow{L} \mathcal{N}(0, I_q(\theta_0)^{-1})$. The rest of the proof is based on the Delta method in the two modes described in Appendix A (Theorems 10 and 11):

(I) $\nabla_\theta \ell(\theta_0; x, y) \neq 0$:

Since the expected log-likelihood function, evaluated at a given pair $(x, y)$, is assumed to be continuously differentiable (A6) and that $\nabla_\theta \ell(\theta_0; x, y) \neq 0$, we can apply Theorem 10 to $\ell(\hat{\theta}_n; x, y) - \ell(\theta_0; x, y)$ to write:

$$\sqrt{n} \left( \ell(\hat{\theta}_n; x, y) - \ell(\theta_0; x, y) \right) \xrightarrow{L} \mathcal{N} \left( 0, \nabla_\theta \ell(\theta_0; x, y) \cdot I_q(\theta_0)^{-1} \cdot \nabla_\theta \ell(\theta_0; x, y) \right),$$ (20)

where the scalar variance can also be written in a trace format.

(II) $\nabla_\theta \ell(\theta_0; x, y) = 0$ and $\nabla_\theta^2 \ell(\theta_0; x, y)$ non-singular:

In this case, the conditions in Theorem 11 are satisfied (with $\Sigma = I_q(\theta_0)^{-1}$ and $g(\theta) = \ell(\theta; x, y)$), and therefore we can directly write (18) and (19) from equations (68) and (69).

Theorem 3 regards the log-likelihood ratio (16) evaluated at any arbitrary pair $(x, y)$. Note that if we consider the training pairs in $L_n$, which are used to obtain $\hat{\theta}_n$, it is known that the ratio evaluated at the training set converges to a single first-degree Chi-square distribution, that is

$$\ell(\hat{\theta}_n; L_n) - \ell(\theta_0; L_n) \xrightarrow{L} \frac{1}{2} \chi^2_1$$ (21)

Also, Theorem 2 implies that variance of the asymptotic distribution of the log-likelihood ratio in case (I) is $\text{tr} \left[ \nabla_\theta \ell(\theta_0; x, y) \cdot \nabla_\theta \ell(\theta_0; x, y) \cdot I_q(\theta_0)^{-1} \right]$, whereas in case (II), from Theorem 11 (see

\[Note that the distribution of the random variable in the right hand side of equation (18) is different from the so-called $Z^2$ distribution (17), since the coefficients $\gamma_{ij}$ can be negative.]
Appendix A, the variance is $\frac{1}{T} \left\| I_q(\theta_0)^{-1/2} \nabla^2 \ell(\theta_0; x, y) I_q(\theta_0)^{-1/2} \right\|^2$. Therefore, it is evident that the variance of the log-likelihood ratio at any $(x, y)$ is reciprocally dependent on the training Fisher information. From this point of view, one can set the training distribution such that it leads to a Fisher information that minimizes this variance. Unless the parameter and hence the Fisher information is univariate, it is not clear what objective to optimize with respect to $q$ such that the resulting Fisher information minimizes the variance. Next, we show that the Fisher information ratio, $\text{tr} \left[ I_q(\theta_0)^{-1} I_p(\theta_0) \right]$, is a reasonable candidate objective to minimize to get a training distribution $q$ for the multivariate case.

**Theorem 4.** If the assumptions $[A0]$ to $[A8]$ hold, then:

$$
\mathbb{E}_p \left[ \text{Var}_q \left( \lim_{n \to \infty} \sqrt{n} \cdot [\ell(\hat{\theta}_n; x, y) - \ell(\theta_0; x, y)] \right) \right] \leq \text{tr} \left[ I_q(\theta_0)^{-1} I_p(\theta_0) \right].
$$

(22)

The approximate equality holds when the set of pairs $(x, y)$ where we have zero score function at $\theta_0$, i.e. $\nabla_\theta \ell(\theta_0; x, y) = 0$, has measure zero under the true joint distribution $P_{\theta_0}$ in $X \times \{1, \ldots, c\}$.

**Proof.** Note that, from Corollary 2 when $\nabla_\theta \ell(\theta_0; x, y) = 0$ the convergence rate of the log-likelihood ratio is one and therefore it is of $O_p \left( \frac{1}{n} \right)$. Thus, in this case we have $\sqrt{n} \cdot [\ell(\hat{\theta}_n; x, y) - \ell(\theta_0; x, y)] = O_p \left( \frac{1}{\sqrt{n}} \right)$ and it converges to zero in probability (and in law). Now, define the region $R_0 \subseteq X \times \{1, \ldots, c\}$ by

$$
R_0 := \{ (x, y) | \nabla_\theta \ell(\theta_0; x, y) = 0 \}
$$

(23)

Variance of the asymptotic distribution of $\sqrt{n} \cdot [\ell(\hat{\theta}_n; x, y) - \ell(\theta_0; x, y)]$, considering both cases $\nabla_\theta \ell(\theta_0; x, y) = 0$ (with probability $P_{\theta_0}(R_0)$) and $\nabla_\theta \ell(\theta_0; x, y) \neq 0$ (with probability $1 - P_{\theta_0}(R_0)$), can be written as:

$$
\text{Var} \left( \lim_{n \to \infty} \sqrt{n} \cdot [\ell(\hat{\theta}_n; x, y) - \ell(\theta_0; x, y)] \right)
\begin{align*}
&= [1 - P_{\theta_0}(R_0)] \cdot \text{tr} \left[ \nabla_\theta \ell(\theta_0; x, y) \cdot \nabla^2 \ell(\theta_0; x, y) \cdot I_q(\theta_0)^{-1} \right] + P_{\theta_0}(R_0) \cdot 0 \\
&\leq \text{tr} \left[ \nabla_\theta \ell(\theta_0; x, y) \cdot \nabla^2 \ell(\theta_0; x, y) \cdot I_q(\theta_0)^{-1} \right]
\end{align*}
$$

(24)

Taking the expectation of both sides with respect to the true joint, gives the inequality (22). If the set of pairs $(x, y)$ where $\nabla_\theta \ell(\theta_0; x, y) = 0$ form a zero measure set under $P_{\theta_0}$, then $P_{\theta_0}(R_0) = 0$ and we get equality in (24) and hence an equality in (22).

Theorem 4 implies that minimizing the Fisher information ratio with respect to $q$, is indeed the upper-bound minimization of the expected variance of the asymptotic distribution of the log-likelihood ratio. However, it cannot be used in practice directly because of the presence of the unknown test distribution $p$ and the true parameter $\theta_0$. We will address this practical issue in details in section 4.3, but here we should mention that part of the dependence on $\theta_0$ may be relaxed using the distribution approximation based on Corollary 1.

**Corollary 2.** If the assumptions $[A0]$ to $[A8]$ hold, then, at any given $(x, y) \in X \times \{1, \ldots, c\}$:

1. In case $\nabla_\theta \ell(\theta_0; x, y) \neq 0$, the likelihood ratio follows an asymptotic normality with convergence rate equal to 1/2. More specifically

$$
\sqrt{n} \cdot [\ell(\hat{\theta}_n; x, y) - \ell(\theta_0; x, y)] \xrightarrow{d} \mathcal{N} \left( 0, \text{tr} \left[ \nabla_\theta \ell(\theta_0; x, y) \cdot \nabla^2 \ell(\theta_0; x, y) \cdot I_q(\theta_0)^{-1} \right] \right).
$$

(25)
In case $\nabla_\theta \ell(\theta_0; x, y) = 0$ and $\nabla^2_\theta \ell(\theta_0; x, y)$ is non-singular, the asymptotic distribution of the log-likelihood ratio is a mixture of first-order Chi-square distributions, and the convergence rate is one. More specifically:

$$n \cdot \left( \ell(\hat{\theta}_n; x, y) - \ell(\theta_0; x, y) \right) \xrightarrow{L} \sum_{i,j=1}^d \left[ \gamma_{ij} \cdot \chi^2_1 \right]$$

where

$$\gamma_{ij} = \left[ I_q(\hat{\theta}_n)^{-1/2} \nabla^2_\theta \ell(\theta_0; x, y) I_q(\hat{\theta}_n)^{-1/2} \right]_{ij}$$

Proof. Follow the same procedure as in the proof of Theorem 3, except that in the beginning use $\sqrt{n} \cdot (\hat{\theta}_n - \theta_0) \xrightarrow{L} N(0, I_q(\hat{\theta}_n)^{-1})$ based on Corollary 1.

Using this corollary to obtain a similar result stated in Theorem 4, one can get $\text{tr} \left[ I_q(\hat{\theta}_n)^{-1} I_p(\theta_0) \right]$ as the asymptotic upper-bound of the expected variance of log-likelihood ratio.

4.2 Application to Active Learning

In this section, we explain how inequality (22) can be utilized as an objective function for active learning. We can perform active learning with this objective by first learning the optimal proposal distribution $q^*$ that minimizes the left-hand-side of inequality (22) and then drawing the queries from this optimal distribution.

In this work, we have shown in Theorem 4 that the Fisher information ratio is an upper-bound for the expected asymptotic variance of the loss function (shown in (16)) of the estimate and the true parameter value. Ideally, one would like to minimize this variance. However, minimizing this expected asymptotic variance (left-hand-side of (22)) is difficult in practice; instead, active learning algorithms [27, 28, 41, 48] minimize the Fisher information ratio (right-hand side of (22)), which is an upper-bound for this expected asymptotic variance.

The inequality (22) is derived under the assumption that $q$ does not change during the sampling process. However, in practice, active sample querying is usually an iterative process. At each iteration, a proposal distribution $q$ is estimated based on the previously observed labels and a finite number of samples is drawn from $q$ as the queries, the labels of which are learned from an oracle. If the learning algorithm requires a larger training data, a new proposal is estimated by taking into account also the newly observed labels. Using inequality (22) to generate such a dynamic process would not be justified as the left-hand-side indicates an asymptotic function of a fixed proposal distribution. Nevertheless, this issue can be resolved by noting that the Fisher information ratio, evaluated at any marginal $q$, is an upper-bound of the global minimum of the original objective. Let the optimal solution be $q^*$:

$$q^* = \arg \min_q J_{p, \theta_0}(q) = \mathbb{E}_p \left[ \text{Var}_q \left( \lim_{n \to \infty} \sqrt{n} \cdot (\ell(\hat{\theta}_n; x, y) - \ell(\theta_0; x, y)) \right) \right]$$

It is evident that for any marginal $q$ we have

$$J_{p, \theta_0}(q^*) \leq J_{p, \theta_0}(q) \leq \text{tr} \left[ I_q(\hat{\theta}_n)^{-1} I_p(\theta_0) \right]$$

Then, by minimizing the Fisher information ratio at each querying iteration with respect to $q$, we approximate $q^*$ through an upper bound minimization.
4.3 Practical Issues

Here, we briefly discuss practical issues with the Fisher information ratio as an objective to be used in active learning of classifiers. These issues consist of evaluation and optimization difficulties.

4.3.1 Evaluation

Note that it is not possible to obtain an exact evaluation of this objective because of two reasons: (1) lack of a closed form for $I_q(\hat{\theta}_n)$ for most of the cases, unless $q$ has a simple parametric form; (2) and more significantly, dependence of $I_p(\theta_0)$ on the unknown parameter $\theta_0$.

Fortunately, the unlabeled input instances are usually easy to obtain. Therefore, Monte Carlo approximation can be employed to calculate $I_q(\hat{\theta}_n)$. Let $X_q$ be the set of samples drawn independently from $q(x)$, then the Fisher information of $q(x)$, at any $\theta \in \Omega$, can be approximated by:

$$I_q(\theta) \approx \hat{I}(\theta; X_q) := \frac{1}{|X_q|} \sum_{i=1}^{c} \sum_{y_i=1}^{|y_i|} p(y|x, \theta) \nabla_\theta \ell(\theta; x_i, y_i) \nabla^\top_\theta \ell(\theta; x_i, y_i) + \delta \cdot I_d$$

where $\delta$ is a small positive number and the weighted identity matrix is added to ensure positive definiteness. If $q(x)$ is either intractable to sample from or not available at all, we can also use the input instances from the previous runs of active learning (if any).

The issue with $I_p(\theta_0)$ is more crucial because of the existence of $\theta_0$. In many approaches \cite{20, 27, 28, 41}, $I_p(\theta_0)$ is replaced by $I_p(\hat{\theta}_n)$. Such replacement is especially useful in pool-based active learning, when there are available test instances generated from $p(x)$ so that a similar Monte-Carlo approximation can be used for $I_p(\hat{\theta}_n)$. More specifically, denote the observed test instances by $X_p$, then we would have the approximation $I_p(\hat{\theta}_n) \approx \hat{I}(\theta_n; X_p)$, yielding the following approximation for the objective:

$$\text{tr} \left[ \hat{I}(\theta_n; X_q)^{-1} \hat{I}(\theta_n; X_p) \right].$$

It is important to remark that when using equation (31), we are actually utilizing some of the test samples in the training process, hence we cannot use those in $X_p$ in order to evaluate the performance of the trained classifier.

4.3.2 Minimization

First note that the matrix $I_p(\theta_0)$ is constant with respect to the new queries. Hence, in the scalar case, we can simply focus on maximizing the training Fisher information. In the multivariate case, though it is not clear what measure of $I_q(\hat{\theta}_n)$ to optimize; one may choose the objective to be $|I_q(\hat{\theta}_n)|$ (where $|\cdot|$ is the determinant function)\footnote{Similar to D-optimality in Optimal Experiment Design \cite{16}.} or $\text{tr}[I_q(\hat{\theta}_n)]$\footnote{Similar to A-optimality in Optimal Experiment Design \cite{16}.} The latter is worth paying more attention due to the following inequality \cite{46}:

$$\text{tr}[I_q(\hat{\theta}_n)^{-1} I_p(\theta_0)] \leq \text{tr}[I_q(\hat{\theta}_n)^{-1}] \cdot \text{tr}[I_p(\theta_0)].$$

(32)
Since \( \text{tr}[I_p(\theta_0)] \) is a constant with respect to \( q \), minimizing the right-hand-side of inequality (22) can itself be approximated by another upper-bound minimization:

\[
\arg \min_q \text{tr}[I_q(\hat{\theta}_n)^{-1}] \tag{33}
\]

A lower bound can also be established for the Fisher information ratio. Using the inequality between arithmetic and geometric means of the eigenvalues of \( I_q(\hat{\theta}_n)^{-1}I_p(\theta_0) \), one can see that

\[
d \cdot |I_q(\hat{\theta}_n)^{-1} - I_p(\theta_0)| \leq \text{tr}[I_q(\hat{\theta}_n)^{-1}I_p(\theta_0)].
\]

Hence, when minimizing the upper-bound by minimizing the trace of \( I_q(\hat{\theta}_n)^{-1} \), one should be careful about the determinant of this matrix as a term influencing the lower-bound of the objective.

In practice, of course, the minimization in (33) can be difficult due to matrix inversion. Thus, sometimes it is further approximated by

\[
\arg \max_q \text{tr}[I_q(\hat{\theta}_n)]. \tag{34}
\]

Hence, algorithms that aim to maximize \( \text{tr}[I_q(\hat{\theta}_n)] \), indeed introduce three layers of objective approximations through equations (32) to (34). In the final objective that only depends on the training Fisher information (either (33) or (34)), \( I_q \) can be replaced by its Monte-Carlo approximation, \( \hat{I}(\cdot; X_q) \), as defined in (30).

However, in pool-based active learning the queries \( X_q \) are to be selected from a pool of unlabeled samples \( X_p \) that are generated from the test marginal \( p(x) \). Having access to a set of samples drawn from the test distribution, enables us to go around the three layers of approximate minimization by approximating both \( I_p \) and \( I_q \) with Monte-Carlo simulations and use equation (31) as the objective.

### 4.4 Algorithms

In this section, we discuss few existing algorithms for implementing the query selection task based on minimization of the Fisher information ratio. We will analyze these algorithms in the context of our unifying framework. These algorithms iterate between selecting the queries and updating the classifier. We assume an initial training set \( \mathcal{L}_{n_0} = \{(x_1, y_1), \ldots, (x_{n_0}, y_{n_0})\} \) is given based on which an initial MLE \( \hat{\theta}_{n_0} \) can be obtained. The initial MLE enables us to approximate the active learning objective function and therefore select queries for building the new training set. After obtaining the query set \( X_q \), for each individual sample \( x \in X_q \), we request its labels \( y(x) \) from the oracle (or equivalently, sample it from the true conditional, \( y(x) \sim p(y|x, \theta_0) \)). These pairs are then added into the training set to get \( \mathcal{L}_{n_1} \), which in turn, is used to update the MLE to \( \hat{\theta}_{n_1} \). Size of the new training data is \( n_1 = n_0 + |X_q| \). This procedure can be done repeatedly for a desirable number of iterations.

In our analysis of the algorithms, we always focus on a specific iteration indexed by \( i \) (as a positive integer). Similar to the notations used above, in this iteration \( \mathcal{L}_{n_{i-1}} \) is the available training data set and \( \theta_{n_{i-1}} \) is the MLE obtained accordingly. For simplicity, we denote the selected query set and the underlying proposal distribution without index \( i \), by \( X_q \) and \( q(x) \), respectively.

The queries can be obtained from two possible optimizations:
• **Probabilistic query selection**: First, optimizing the objective with respect to the proposal distribution \(q(x)\), and then generate the queries independently from the obtained proposal \((X_q \sim q(x))\);

• **Deterministic query selection**: Optimizing the objective directly with respect to the query set \(X_q\).

In probabilistic active learning, with a given MLE, the queries \(X_q\) cannot be selected deterministically. Whereas in deterministic active learning, once the MLE is obtained, there is no randomness in selecting \(X_q\), discarding the uncertainty in the employed numerical optimization techniques.

Furthermore, it is also useful to divide the querying algorithms into two categories based on the size of \(X_q\): **sequential active learning** where a single sample is queried at each iteration, i.e. \(|X_q| = 1\); and **batch active learning** where the size of the query set is larger than one. The non-singleton query batches are usually generated greedily, with the batch size \(|X_q|\) fixed to a constant value.

In the remainder of this paper, we discuss four algorithms that somehow used Fisher information ratio as their active learning objective: one probabilistic algorithm for batch active learning, followed by one algorithm for sequential deterministic querying and two algorithms for deterministic batch active learning.

### 4.4.1 Probabilistic Active Learning

In this section, we give a simple probabilistic algorithm for sequential querying within the framework described above. This algorithm is the classification version of the probabilistic active learning proposed by Fukumizu [20] for regression problem.

**Algorithm 1 (Fukumizu [20])**

The assumption is that the proposal belongs to a parametric family and is of the form \(q(x; \alpha)\), where \(\alpha\) is the parameter vector of the family. In this parametric active learning, at iteration \(i\), the best set of parameters \(\alpha_i\) is selected based on a previously constructed training set \(L_{n_i-1}\), and the query set is sampled from the resulting proposal distribution \(X_q \sim q(x; \alpha_i)\).

Since there is no access to the test samples in this algorithm, it uses the simplified optimization in its formulation. The new parameter vector is obtained by maximizing the expected contribution of the queries generated from \(q(x; \alpha)\) to this objective. In order to give a Monte-Carlo approximation of the Fisher information matrix, we use all the input instances we already selected in previous iterations, as well as the new queries. More specifically, denoting the covariates of \(L_{n_i-1}\) by \(X_{i-1}\), we use \(I(\hat{\theta}_{n_i-1}; X_{i-1})\) and \(I(\hat{\theta}_{n_i-1}; X_{i-1} \cup X_q)\) as Fisher information approximations before and after selecting the queries, respectively. The parameter vector \(\alpha_i\) is selected to maximize the expected trace value of the latter approximation, that is

\[
E_q(x; \alpha) \left[ \text{tr} \left[ I(\hat{\theta}_{n_i-1}; X_{i-1} \cup X_q) \right] \right] = \text{tr} \left[ \frac{n_{i-1}}{n_i} I(\hat{\theta}_{n_i-1}; X_{i-1}) + \frac{1}{n_i} E_q(x; \alpha) [I(\hat{\theta}_{n_i-1}; X_q)] \right] .
\]  

(35)

The first term is independent of the query set \(X_q\), hence we focus only on the second term in our
optimization. Noting that the queries are generated independently, we can rewrite this term as:

\[
\mathbb{E}_{q(x; \alpha_i)} \left[ \text{tr} \left[ \hat{I}(\hat{\theta}_{n_{i-1}}; x) \right] \right] = \frac{1}{|X_q|} \sum_{x \in X_q} \mathbb{E}_{q(x; \alpha_i)} \left[ \text{tr} \left[ \hat{I}(\hat{\theta}_{n_{i-1}}; x) \right] \right] - (|X_q| - 1)\delta \cdot I_d
\]

Equation (36) implies that the parameter vector \( \alpha_i \) can be obtained by maximizing the expected contribution of a single query to the Fisher information, that is:

\[
\alpha_i = \arg \max_{\alpha} \mathbb{E}_{q(x; \alpha_i)} \left[ \text{tr} \left[ \hat{I}(\hat{\theta}_{n_{i-1}}; x) \right] \right] - (|X_q| - 1)\delta \cdot I_d
\]

The procedure is shown in Algorithm 1. It can be used in both sequential and batch modes by changing the number of samples drawn from \( q(x; \alpha_i) \).

We emphasize that Algorithm 1 is probabilistic, meaning that at any iteration \( i \), with fixed \( \hat{\theta}_{n_{i-1}} \) and \( X_{n_{i-1}} \), the next set of queries are not deterministically selected. The optimization is performed with respect to the parameters of the proposal distribution, which are then used to sample \( X_q \). Fukumizu [20] claims that introducing such randomness into active learning, which increases exploration against exploitation, may prevent the algorithm from falling into local optima. Also note that this algorithm is not pool-based, meaning that it does not select the queries from a pool of observed instances, although could be constrained to do so.

4.4.2 Deterministic Active Learning

Here, we analyze three deterministic algorithms previously developed for querying based on Fisher information ratio. The first algorithm of such kind that is proposed for active learning in classification, has been originally proposed by Zhang and Oles [48]. The other two algorithms we discuss here, which are specific for pool-based active learning, are both inspired by this original work.

**Algorithm 2 (Zhang and Oles [48])**

Zhang and Oles [48] started from optimization problem (34), and introduced additional simplifications to it, specifically considering the use of a binary logistic regression classifier. Here, we discuss their formulation using a general discriminate framework.

In their algorithm, a single query is selected at each iteration. Denote it by \( X_q = \{ \tilde{x} \} \) and consider iteration \( i \) (hence fixing \( L_{n_{i-1}}, X_{i-1} \) and \( \hat{\theta}_{n_{i-1}} \)). Similar to the previous section, the Fisher information matrix \( I_q \) can be approximated as in (30). Zhang and Oles [48] discarded the expectation with respect to the proposal distribution in (37) or equivalently consider \( q \) to be a uniform distribution. Therefore, the optimization with respect to parameters turned into a direct optimization with respect to the single query \( \tilde{x} \):

\[
\tilde{x} = \arg \max_{x \in X} \sum_{y=1}^{c} p(y|x, \hat{\theta}_{n_{i-1}}) \nabla_{\theta}^\top \ell(\hat{\theta}_{n_{i-1}}; x, y) \nabla_{\theta} \ell(\hat{\theta}_{n_{i-1}}; x, y)
\]
Algorithm 1: Parametric Active Learning

**Inputs:** The initial training set $\mathcal{L}_{n_0}$; number of querying iterations $i_{max}$

**Outputs:** The trained classifier with MLE $\hat{\theta}_{n_{i_{max}}}$

```plaintext
/* Initializations */
1 $\hat{\theta}_{n_0} \leftarrow \arg\max_{\theta} \ell(\theta; \mathcal{L}_{n_0})$

/* Starting the Iterations */
2 for $i = 1 \rightarrow i_{max}$ do

/* Parameter Optimization */
3 $\alpha_i = \arg\max_{\alpha} \mathbb{E}_{q(x; \alpha)} \left[ \sum_{y=1}^c p(y|x, \hat{\theta}_{n_{i-1}}) \nabla^\top \theta \ell(\hat{\theta}_{n_{i-1}}; x, y) \nabla \theta \ell(\hat{\theta}_{n_{i-1}}; x, y) \right]$

/* Sample from the parametric proposal */
4 $X_q \sim q(x; \alpha_i)$

/* Request the queries' labels from the oracle */
5 $y(x) \sim p(y|x, \theta_0) \ \forall x \in X_q$

/* Taking care of indexing */
6 $n_i \leftarrow n_{i-1} + |X_q|$

/* Update the training set and update MLE */
7 $\mathcal{L}_{n_i} \leftarrow \mathcal{L}_{n_{i-1}} \cup \left\{ \bigcup_{x \in X_q} (x, y(x)) \right\}$
8 $\hat{\theta}_{n_i} \leftarrow \arg\max_{\theta} \ell(\theta; \mathcal{L}_{n_i})$

9 return $\hat{\theta}_{n_{i_{max}}}$
```

This deterministic approach, shown in Algorithm 2, is very similar to the probabilistic approach described above, except that there is no intermediate parameter optimization step.

It is important to note that Algorithm 2 can be used in pool-based active learning as well. This can be done by constraining $x$ to be a member of a pool of samples. It can even be generalized to batch pool active learning by solving such constrained optimization problem multiple times. However, such iterative optimization is not efficient, because the resulting queries will most probably be close to each other and therefore contain redundant information.

Algorithm 3 (Settles and Craven [41])

Inspired by Zhang and Oles [48], Settles and Craven [41] employed Fisher information ratio to develop a pool-based active learning, which can be used in either sequential or batch querying. Suppose that, at iteration $i$, we have access to a pool of unlabeled samples $X_p$, which is assumed to be drawn from the test marginal $p(x)$. The queries are chosen from this pool, that is $X_q \subseteq X_p$. The test Fisher information matrix can be approximated by Monte-Carlo simulation over the samples in $X_p$, meaning $\hat{I}(\hat{\theta}_{n_{i-1}}; X_p)$. Furthermore, as before, we denote the input instance of $\mathcal{L}_{n_{i-1}}$ selected so far, by $X_{i-1}$. As explained in section 4.4.1 the training Fisher information matrix after querying a set $X_q$ can be approximated by $\hat{I}(\hat{\theta}_{n_{i-1}}; X_{i-q} \cup X_q)$. Thus, since we do have an approximation of both Fisher information matrices, the objective to minimize is chosen to be in the form of (31).

Similar to the Zhang and Oles [48] algorithm, the proposal distribution $q$ is ignored in the
Algorithm 2: Deterministic Active Learning of Zhang and Oles [48]

Inputs: The initial training set $L_{n_0}$, number of querying iterations $i_{\text{max}}$

Outputs: The trained classifier with MLE $\hat{\theta}_{n_{\text{max}}}$

/* Initializations */
1 $\hat{\theta}_{n_0} \leftarrow \text{arg max}_\theta \ell(\theta; L_{n_0})$

/* Starting the Iterations */
2 for $i = 1 \rightarrow i_{\text{max}}$ do

/* Query optimization */
3 $X_q = \{\tilde{x}\} = \text{arg max}_x \sum_{y=1}^c p(y|x, \hat{\theta}_{n_{i-1}}) \nabla_\theta^\top \ell(\hat{\theta}_{n_{i-1}}; x, y) \nabla_\theta \ell(\hat{\theta}_{n_{i-1}}; x, y)$

/* Request the query’s label from the oracle */
4 $y(\tilde{x}) \sim p(y|\tilde{x}, \theta_0)$

/* Taking care of indexing */
5 $n_i \leftarrow n_{i-1} + 1$

/* Update the training set and update MLE */
6 $L_{n_i} \leftarrow L_{n_{i-1}} \cup \{(\tilde{x}, y(\tilde{x}))\}$
7 $\hat{\theta}_{n_i} \leftarrow \text{arg max}_\theta \ell(\theta; L_{n_i})$

8 return $\hat{\theta}_{n_{\text{max}}}$

objective (or equivalently considered as being uniform). An additional assumption Settles and Craven [41] made to simplify the optimization task is:

$$\text{arg min}_{X_q \subset X_p} \text{tr} \left[ \tilde{I}(\hat{\theta}_{n_{i-1}}; X_{i-1} \cup X_q)^{-1} \tilde{I}(\hat{\theta}_{n_{i-1}}; X_p) \right] \approx \text{arg min}_{X_q \subset X_p} \text{tr} \left[ \tilde{I}(\hat{\theta}_{n_{i-1}}; X_q)^{-1} \tilde{I}(\hat{\theta}_{n_{i-1}}; X_p) \right].$$

(39)

This simplified optimization is easy to implement for sequential active learning. However, the combinatorial optimization required for batch active learning can easily become intractable. As shown in Algorithm 3, Settles and Craven [41] used a greedy approach to do this optimization (the inner loop).

Algorithm 4 (Hoi et al. [27] and Hoi et al. [28])

The algorithms proposed by Hoi et al. [27] and Hoi et al. [28] are very similar to the one developed by Settles and Craven [41], described above, except that they use a more sophisticated optimization method. Their method shown in Algorithm 4 is different from Algorithm 3 mainly in the way that it greedily chooses the query at each iteration. While Algorithm 3 exclusively considers the contribution of each $x \in X_q$, ignoring the samples selected in the previous iterations (hence $\tilde{I}(\hat{\theta}_{n_i}; x)$ in line 5), Algorithm 4 takes into account all the queries chosen so far (hence $\tilde{I}(\hat{\theta}_{n_i}; X_q \cup \{x\})$ in line 5).

Hoi et al. [27] and Hoi et al. [28] showed that when using binary logistic regression classifier, their optimization (39) can be done by maximizing a submodular set function with respect
Algorithm 3: Deterministic Active Learning of Settles and Craven

**Inputs:** The initial sample pool $X_p$, the initial training set $L_{n_0}$, size of the batches $|X_q|$, number of querying iterations $i_{max}$

**Outputs:** The trained classifier with MLE $\hat{\theta}_{n_{i_{max}}}$

```c
/* Initializations */
1  $\hat{\theta}_{n_0} \leftarrow \arg \max_{\theta} \ell(\theta; L_{n_0})$
/* Starting the Iterations */
2  for $i = 1 \rightarrow i_{max}$ do
    /* Initializing the query set for this iteration */
    3  $X_q \leftarrow \emptyset$
    /* The loop for greedy batch querying */
    4  for $j = 1 \rightarrow |X_q|$ do
        /* Query optimization and adding the result into the query set */
        5  $X_q = X_q \cup \{ \arg \min_{x \in X_p} \text{tr}[\mathbf{I}(\hat{\theta}_{n_{i-1}}; x) - 1 \mathbf{I}(\hat{\theta}_{n_{i-1}}; X_p)] \}$
        /* Removing the selected queries from the pool */
        6  $X_p \leftarrow X_p - X_q$
    /* Request the queries' labels from the oracle */
    7  $y(x) \sim p(y|x, \theta_0) \quad \forall x \in X_q$
    /* Taking care of indexing */
    8  $n_i \leftarrow n_{i-1} + |X_q|$
    /* Update the training set and update MLE */
    9  $L_{n_i} \leftarrow L_{n_{i-1}} \cup \{ \cup_{x \in X_q} (x, y(x)) \}$
  10  $\hat{\theta}_{n_i} \leftarrow \arg \max_{\theta} \ell(\theta; L_{n_i})$
11  return $\hat{\theta}_{n_{i_{max}}}$
```

to the query set $X_q$. This allowed them to use the well-known iterative algorithm proposed by Nemhauser et al. [35], which guarantees a tight lower-bound for maximization of submodular and monotone set functions.

In the rest of this section, we show that minimizing this objective obtained from the above-mentioned assumptions, can be efficiently approximated by a monotonically submodular maximizing under any discriminative classifier. This is a generalization of the result derived by Hoi et al. [27] that is obtained in case of using logistic regression classifier. As a consequence, FIR can be efficiently optimized with guaranteed tight bounds [34, 35]. As the first step, we prove that [39] is approximately equivalent to maximizing a simplified set function, for any unlabeled sample pool $X_p$:

**Lemma 1.** Let $X_p, X_q \subseteq X$ be two non-empty and finite subsets of samples randomly generated from $p(x)$ and its resample distribution $q(x)$, respectively, such that $X_q \subseteq X_p$, and the parameter $\delta \geq 0$ in (30) is a small constant. If assumptions (A0), (A4), (A6) and (A8) hold, then the following optimization problems are approximately equivalent for some function $g_\theta : X \times \{1, \ldots, c\} \times X \rightarrow \mathbb{R}^+$, $d$-dimensional non-zero vector $v_\theta$ depending on $x$ and $y$, and for
Algorithm 4: Deterministic Active Learning of Hoi et al. [27] and Hoi et al. [28]

Inputs: The initial sample pool $X_p$, the initial training set $L_{n_0}$, size of the batches $|X_q|$, number of querying iterations $i_{\text{max}}$

Outputs: The trained classifier with MLE $\hat{\theta}_{n_{\text{imax}}}$

/* Initializations */
1 $\hat{\theta}_{n_0} \leftarrow \arg\max_\theta \ell(\theta; L_{n_0})$
2 /* Starting the Iterations */
3 for $i = 1 \rightarrow i_{\text{max}}$ do
4   /* Initializing the query set */
5     $X_q \leftarrow \emptyset$
6     /* The loop for greedy batch querying */
7     for $j = 1 \rightarrow |X_q|$ do
8       $\hat{x} = \arg\min_{\mathbf{x} \in X_p} \text{tr} \left[ \hat{I}_{\hat{\theta}_{n_{i-1}}}^{-1}(\hat{\theta}_{n_{i-1}}; X_q) - \hat{I}_{\hat{\theta}_{n_{i-1}}}^{-1}(\hat{\theta}_{n_{i-1}}; X_p) \right]$
9       $X_q \leftarrow X_q \cup \{\hat{x}\}$
10      /* Remove the selected instance from the pool */
11      $X_p \leftarrow X_p - \{\hat{x}\}$
12      /* Request the queries’ labels from the oracle */
13      $y(\mathbf{x}) \sim p(y|\mathbf{x}, \theta_0) \quad \forall \mathbf{x} \in X_q$
14      /* Taking care of indexing */
15      $n_i \leftarrow n_{i-1} + |X_q|$
16      /* Update the training set and update MLE */
17      $L_{n_i} \leftarrow L_{n_{i-1}} \cup \{\mathbf{x}, y(\mathbf{x})\}$
18      $\hat{\theta}_{n_i} \leftarrow \arg\max_\theta \ell(\theta; L_{n_i})$
19   return $\hat{\theta}_{n_{\text{imax}}}$

all $\theta \in \Omega$:

(i) $\arg\min_{X_q \subset X_p} \text{tr} \left[ \hat{I}(\theta; X_q)^{-1} \hat{I}(\theta; X_p) \right] \quad (40a)$

(ii) $\arg\max_{X_q \subset X_p} \sum_{\mathbf{x} \in X_p - X_q} \sum_{y=1}^c \delta \cdot \|v_\theta(\mathbf{x}, y)\|^{-2} - \sum_{\mathbf{x}' \in X_q} g_\theta(\mathbf{x}, y, \mathbf{x}') \quad (40b)$

Proof. We first substitute the score function of the classifier

$\nabla_\theta \log p(y|\mathbf{x}, \theta) = \frac{\nabla_\theta p(y|\mathbf{x}, \theta)}{p(y|\mathbf{x}, \theta)}$
Applying the trace function to both sides of the equation will result:

\[
\hat{I}(\theta; X_q) = \frac{1}{|X_q|} \sum_{x \in X_q} \sum_{y=1}^c p(y|x, \theta) \cdot \frac{\nabla_{\theta} p(y|x, \theta) \nabla_{\theta}^\top p(y|x, \theta)}{p(y|x, \theta)^2} + \delta \Pi_d
\]  

(41)

\[
= \frac{1}{|X_q|} \sum_{x \in X_q} \sum_{y=1}^c \frac{\nabla_{\theta} p(y|x, \theta) \nabla_{\theta}^\top p(y|x, \theta)}{p(y|x, \theta)} + \delta \Pi_d
\]  

(42)

Define the vector \( v_\theta(x, y) := \nabla_{\theta} p(y|x, \theta) / \sqrt{p(y|x, \theta)} \) and rewrite \( \hat{I}(\theta; X_q) \) as:

\[
\hat{I}(\theta; X_q) = \frac{1}{|X_q|} \sum_{x \in X_q} \sum_{y=1}^c v_\theta(x, y) \cdot v_\theta(x, y)^\top + \delta \cdot \Pi_d.
\]  

(43)

On the other hand, since \( X_q \subset X_p \) we can write \( \hat{I}(\theta; X_p) \) in terms of \( \hat{I}(\theta; X_q) \) by breaking the summation over \( X_p \) into summations over \( X_q \) and \( X_p - X_q \) as follows:

\[
\hat{I}(\theta; X_p) = \left( \frac{|X_q|}{|X_p|} \right) \cdot \hat{I}(\theta; X_q) + \frac{1}{|X_p|} \sum_{x \in X_p - X_q} \sum_{y=1}^c v_\theta(x, y) \cdot v_\theta(x, y)^\top + \delta \left( \frac{|X_p| - |X_q|}{|X_p|} \right) \cdot \Pi_d
\]  

(44)

Now that we related the Fisher information matrices to each other, we can compute the product of \( \hat{I}(\theta; X_p) \) and \( \hat{I}(\theta; X_q)^{-1} \):

\[
\hat{I}(\theta; X_q)^{-1} \hat{I}(\theta; X_p) = \left( \frac{|X_q|}{|X_p|} \right) \cdot \Pi_d + \frac{\hat{I}(\theta; X_q)^{-1}}{|X_p|} \left[ \sum_{x \in X_p - X_q} \sum_{y=1}^c v_\theta(x, y) \cdot v_\theta(x, y)^\top \right] + \delta \left( \frac{|X_p| - |X_q|}{|X_p|} \right) \cdot \hat{I}(\theta; X_q)^{-1}
\]  

(45)

Applying the trace function to both sides of the equation will result:

\[
\text{tr} \left[ \hat{I}(\theta; X_q)^{-1} \hat{I}(\theta; X_p) \right] = \frac{|X_q|}{|X_p|} \cdot d + \frac{1}{|X_p|} \sum_{x \in X_p - X_q} \sum_{y=1}^c \text{tr} \left[ \hat{I}(\theta; X_q)^{-1} v_\theta(x, y) \cdot v_\theta(x, y)^\top \right] + \delta \left( \frac{|X_p| - |X_q|}{|X_p|} \right) \cdot \text{tr} \left[ \hat{I}(\theta; X_q)^{-1} \right]
\]

\[
\approx \frac{|X_q|}{|X_p|} \cdot d + \frac{1}{|X_p|} \sum_{x \in X_p - X_q} \sum_{y=1}^c v_\theta(x, y)^\top \hat{I}(\theta; X_q)^{-1} v_\theta(x, y),
\]  

(46)
where the last term is dropped since the overloading constant, $\delta$, is assumed to be small. Furthermore, the term including $\hat{I}(\theta; X_q)^{-1}$ can be approximated by replacing the weighted harmonic mean of the eigenvalues of $\hat{I}(\theta; X_q)$ by their weighted arithmetic mean [27]:

$$v_\theta(x, y)^\top \hat{I}(\theta; X_q)^{-1} v_\theta(x, y) \approx \frac{\|v_\theta(x, y)\|^4}{v_\theta(x, y)^\top \hat{I}(\theta; X_q) v_\theta(x, y)}. \quad (47)$$

Note that this approximation becomes exact when the condition number of $\hat{I}(\theta; X_q)$ is one. Substituting $\hat{I}(\theta; X_q)$ from equation (43) into the denominator of the approximation above yields:

$$v_\theta(x, y)^\top \hat{I}(\theta; X_q) v_\theta(x, y) = \frac{1}{|X_q|} \sum_{x' \in X_q} \sum_{y=1}^c \left[ v_\theta(x, y)^\top v_\theta(x', y') \right]^2 + \delta \|v_\theta(x, y)\|^2 \quad (48)$$

Integrating this approximation with equation (46), and assuming that the value of $\theta$ is not located at the stationary point of the conditional density $p(y|x, \theta)$ (hence $v_\theta(x, y)$ is not the zero vector), results:

$$\text{tr} \left[ \hat{I}(\theta; X_q)^{-1} \hat{I}(\theta; X_p) \right] \approx \frac{|X_q| \cdot d}{|X_p|} + \frac{1}{|X_p|} \sum_{x \in X_p - X_q} \sum_{y=1}^c \delta \cdot \|v_\theta(x, y)\|^{-2} + \frac{1}{|X_q|} \sum_{x' \in X_q} g_\theta(x, y, x') \quad (49)$$

where

$$g_\theta(x, y, x') := \frac{1}{|X_q|} \sum_{y'=1}^c \left[ v_\theta(x, y)^\top v_\theta(x', y') \right]^2 \|v_\theta(x, y)\|^2 \quad (50)$$

Finally, in (49), removing the constants we get

$$\arg \min_{X_q \subseteq X_p} \text{tr} \left[ \hat{I}(\theta; X_q)^{-1} \hat{I}(\theta; X_p) \right] \approx \arg \max_{X_q \subseteq X_p} \sum_{x \in X_p - X_q} \sum_{y=1}^c \delta \cdot \|v_\theta(x, y)\|^{-2} + \frac{1}{|X_q|} \sum_{x' \in X_q} g_\theta(x, y, x') \quad (51)$$

Note that Lemma 1, as stated above, does not depend on the size of $X_q$. However, just as before, in practice it is usually assumed that $|X_q| > 0$ is fixed and therefore the optimizations in (41) should be considered with cardinality constraint. In general, combinatorial maximization problems can turn out to be intractable. Next, it is shown that the objective at hand is a monotonically submodular set function in terms of $X_q$ and therefore can be maximized efficiently with a greedy approach such as that shown in Algorithm 4.

**Theorem 5.** Suppose $f_\theta : 2^{X_p} \to \mathbb{R}$ is defined as:

$$f_\theta(X_q) = \sum_{x \in X_p - X_q} \sum_{y=1}^c \delta \cdot \|v_\theta(x, y)\|^{-2} + \frac{-1}{|X_q|} \sum_{x' \in X_q} g_\theta(x, y, x'), \quad \forall X_q \subseteq X_p \quad (52)$$

with $v_\theta$ a $d$-dimensional vector depending on $x$ and $y$, and $g_\theta$ defined in (47). Then $f_\theta$ is a submodular and monotone (non-decreasing) set function for all $\theta \in \Omega$. 

22
Proof. First, note that the function $f_\theta$ can be broken into simpler terms $f_\theta(X_q) = \sum_{y=1}^{c} f_\theta(X_q; y)$, where

$$f_\theta(X_q; y) = \sum_{x \in X_p - X_q} \frac{-1}{\delta \cdot \|v_\theta(x, y)\|^2 + \sum_{x' \in X_q} g_\theta(x, y, x')}, \quad \forall X_q \subseteq X_p. \quad (53)$$

Therefore, in order to prove submodularity and monotonicity of $f_\theta$, it suffices to prove these properties for $f_\theta(\cdot; y)$ for all $y \in \{1, \ldots, c\}$. Fix $y$ and take any subset $X_q \subseteq X_p$ and $\xi \in X_p - X_q$. Then, we can write:

$$f_\theta(X_q \cup \{\xi\}; y) = \sum_{x \in X_p - X_q \cup \{\xi\}} \frac{-1}{\delta \cdot \|v_\theta(x, y)\|^2 + \sum_{x' \in X_q \cup \{\xi\}} g_\theta(x, y, x')} \quad (54)$$

$$= \sum_{x \in X_p - X_q} \frac{-1}{\delta \cdot \|v_\theta(x, y)\|^2 + \sum_{x' \in X_q} g_\theta(x, y, x')} + \frac{1}{\delta \cdot \|v_\theta(\xi, y)\|^2 + \sum_{x' \in X_q \cup \{\xi\}} g_\theta(\xi, y, x')}.$$

We then form the discrete derivative of $f_\theta(\cdot; y)$ at $X_q$ to get:

$$\rho_{f_\theta(\cdot; y)}(X_q; \xi) = f_\theta(X_q \cup \{\xi\}; y) - f_\theta(X_q; y)$$

$$= \sum_{x \in X_p - X_q} \left[ \frac{-1}{\|v_\theta(x, y)\|^2 + \sum_{x' \in X_q \cup \{\xi\}} g_\theta(x, y, x')} + \frac{1}{\|v_\theta(\xi, y)\|^2 + \sum_{x' \in X_q \cup \{\xi\}} g_\theta(\xi, y, x')} \right]$$

$$+ \frac{\delta}{\|v_\theta(x, y)\|^2} + \frac{1}{\|v_\theta(\xi, y)\|^2}.$$  (55)

The right-hand-side can be rewritten as

$$\sum_{x \in X_p - X_q} \left[ \frac{g_\theta(x, y, \xi)}{\delta \cdot \|v_\theta(x, y)\|^2 + \sum_{x' \in X_q \cup \{\xi\}} g_\theta(x, y, x')} \right]$$

$$+ \frac{1}{\|v_\theta(x, y)\|^2} + \frac{1}{\|v_\theta(\xi, y)\|^2}. \quad (56)$$

Since by definition $g_\theta(x, y, x') \geq 0, \forall x, y, x'$, all of the terms in (56) are non-negative and therefore $\rho_{f_\theta(\cdot; y)}(X_q; \xi) \geq 0$. This is true for any $X_q \subseteq X_p$ hence monotonicity of $f_\theta(\cdot; y)$ is obtained. Now let us take any superset $X_{q'}$ such that $X_q \subseteq X_{q'} \subseteq X_p$ and $\xi \in X_p - X_{q'}$, and form the
difference between their corresponding discrete derivatives. From (56) we will have:

$$\rho_{f_{\theta}(\cdot; y)}(X_q; \xi) - \rho_{f_{\theta}(\cdot; y)}(X_{q'}; \xi)$$

$$= \sum_{x \in X_p - X_q} \left[ \frac{g_\theta(x, y, \xi)}{\|v_\theta(x, y)\|^2} + \sum_{x' \in X_q \cup \{\xi\}} g_\theta(x, y, x') \right] \left( \frac{\delta}{\|v_\theta(x, y)\|^2} + \sum_{x' \in X_q} g_\theta(x, y, x') \right)$$

$$+ \frac{\delta}{\|v_\theta(x, y)\|^2} + \sum_{x' \in X_q \cup \{\xi\}} g_\theta(x, \xi, x')$$

$$- \sum_{x' \in X_{q'} - X_q} \left[ \frac{g_\theta(x, y, x')}{\|v_\theta(x, y)\|^2} + \sum_{x' \in X_q \cup \{\xi\}} g_\theta(x, y, x') \right] \left( \frac{\delta}{\|v_\theta(x, y)\|^2} + \sum_{x' \in X_q} g_\theta(x, y, x') \right)$$

$$- \frac{\delta}{\|v_\theta(x, y)\|^2} - \sum_{x' \in X_q \cup \{\xi\}} g_\theta(x, \xi, x') \quad (57)$$

From non-negativity of $g_\theta$ and that $X_q \subseteq X_{q'}$, we can conclude that for any $x \in X$ and $y \in \{1, ..., c\}$:

$$\sum_{x' \in X_{q'}} g_\theta(x, y, x') \geq \sum_{x' \in X_q} g_\theta(x, y, x')$$

$$\Leftrightarrow \left[ \sum_{x' \in X_{q'}} g_\theta(x, y, x') + \frac{\delta}{\|v_\theta(x, y)\|^2} \right]^{-1} \leq \left[ \sum_{x' \in X_q} g_\theta(x, y, x') + \frac{\delta}{\|v_\theta(x, y)\|^2} \right]^{-1}$$

$$\Leftrightarrow -\left[ \sum_{x' \in X_{q'} \cup \{\xi\}} g_\theta(x, y, x') + \frac{\delta}{\|v_\theta(x, y)\|^2} \right]^{-1} \geq -\left[ \sum_{x' \in X_q \cup \{\xi\}} g_\theta(x, y, x') + \frac{\delta}{\|v_\theta(x, y)\|^2} \right]^{-1} \quad (58)$$

Similarly, since $X_q \cup \{\xi\} \subseteq X_{q'} \cup \{\xi\}$ we will get:

$$-\left[ \sum_{x' \in X_{q'} \cup \{\xi\}} g_\theta(x, y, x') + \frac{\delta}{\|v_\theta(x, y)\|^2} \right]^{-1} \geq -\left[ \sum_{x' \in X_q \cup \{\xi\}} g_\theta(x, y, x') + \frac{\delta}{\|v_\theta(x, y)\|^2} \right]^{-1} \quad (59)$$

Applying the inequalities (58) and (59) into equation (57) results:

$$\rho_{f_{\theta}(\cdot; y)}(X_q; \xi) - \rho_{f_{\theta}(\cdot; y)}(X_{q'}; \xi)$$

$$\geq \sum_{x \in X_p - X_q} \left[ \frac{g_\theta(x, y, \xi)}{\|v_\theta(x, y)\|^2} + \sum_{x' \in X_q \cup \{\xi\}} g_\theta(x, y, x') \right] \left( \frac{\delta}{\|v_\theta(x, y)\|^2} + \sum_{x' \in X_q} g_\theta(x, y, x') \right)$$

$$- \sum_{x \in X_p - X_{q'}} \left[ \frac{g_\theta(x, y, \xi)}{\|v_\theta(x, y)\|^2} + \sum_{x' \in X_q \cup \{\xi\}} g_\theta(x, y, x') \right] \left( \frac{\delta}{\|v_\theta(x, y)\|^2} + \sum_{x' \in X_q} g_\theta(x, y, x') \right)$$

$$+ \frac{\delta}{\|v_\theta(x, y)\|^2} + \sum_{x' \in X_q \cup \{\xi\}} g_\theta(x, \xi, x') - \frac{\delta}{\|v_\theta(x, y)\|^2} + \sum_{x' \in X_q \cup \{\xi\}} g_\theta(x, \xi, x') \quad (60)$$
which yields

\[
\sum_{x \in X' \setminus X_q} g_\theta(x, y, \xi) \left( \frac{\delta}{\|v_\theta(x,y)\|^2} + \sum_{x' \in X_q \cup \{\xi\}} g_\theta(x, y, x') \right) \geq 0. \tag{61}
\]

Inequality (61) holds for any \(X_q \subseteq X_p\); hence submodularity of \(f_\theta(\cdot; y)\) stands for all \(y \in \{1, \ldots, c\}\) and \(\theta \in \Omega\).

The result above, together with Lemma 1, imply that the objective of (40b) is a monotonically increasing set function with respect to \(X_q\). Below we present the main result that guarantees tight bounds for greedy maximization of monotonic submodular set functions. Details of this result, which is also shown to be the optimally efficient solution to submodular maximization, can be found in the seminal papers by Nemhauser et al. [35] and Nemhauser and Wolsey [34].

**Theorem 6** (Nemhauser et al. [35]). Let \(f_\theta : 2^{X_p} \to \mathbb{R}\) be any submodular and nondecreasing set function with \(f(\emptyset) = 0\). If \(X_q\) is the output of a greedy maximization algorithm, and \(X_q^*\) is the optimal maximizer of \(f_\theta\) with a cardinality constraint (fixed \(|X_q|\)), then we have:

\[
f_\theta(X_q) \geq \left[ 1 - \left( \frac{|X_q| - 1}{|X_q|} \right)^{|X_q|} \right] f_\theta(X_q^*) \geq \left( 1 - \frac{1}{e} \right) f_\theta(X_q^*). \tag{62}
\]

In Algorithm 4, the inner loop (lines 4 to 7) implements the minimization in (39) greedily. We have seen above that this set minimization is equivalent to maximizing a submodular and monotone set maximization, which, in turn, is shown to be optimally efficient.

### 5 Conclusion

In this paper, we focused on active learning algorithms in classification problems whose objectives are based on Fisher information criterion. As the primary result, we showed the dependency of the variance of the asymptotic distribution of log-likelihood ratio on the Fisher information of the training distribution. Then, we used this dependency to derive our novel theoretical contribution by establishing the Fisher information ratio (FIR) as the upper bound of such asymptotic variance. We discussed that several layers of approximations can be employed in practice to simplify FIR; simplifications, that can usually be avoided in pool-based active learning. Finally Monte-Carlo simulations and greedy algorithms can be used to evaluate and optimize the (simplified) FIR objective, respectively. Using this framework, we can distinguish the main differences between some of the FIR-based querying methods in the classification context. Such comparative analysis, not only shed light on the assumptions and simplifications of the existing algorithms, it
can also be helpful for finding suitable directions in developing novel active learning algorithms based on the Fisher information criterion.

Appendices

A  Statistical Background

Asymptotic analysis plays an important role in statistics. It considers the extreme cases where the number of observations is increased with no bounds. In such scenarios, discussions on different notions of convergence of the sequence of random variables naturally arise. Generally speaking, there are three major types of stochastic convergence: convergence in probability, convergence in law (distribution) and convergence with high probability (almost surely). Here, we focus on the two former modes of convergence, discuss two fundamental results based on them and formalize our notations regarding parameter estimators. Further details of the following definitions and results can be found in any standard statistical textbook such as Lehmann and Casella [31].

A.1 Convergence of Random Variables

Throughout this section, \( \{ \theta_1, \theta_2, ..., \theta_n, ... \} \), denoted simply by \( \{ \theta_n \} \), is a sequence of multivariate random variables lying in \( \Omega \subseteq \mathbb{R}^d \). Also suppose that \( \theta_0 \) is a constant vector and \( \tilde{\theta} \) is another random variable in the same space \( \Omega \).

Definition 1. We say that the sequence \( \{ \theta_n \} \) converges in probability to \( \theta_0 \) and write \( \theta_n \overset{P}{\rightarrow} \theta_0 \), iff for every \( \varepsilon > 0 \) we have:

\[
P(|\theta_{ni} - \theta_{0i}| > \varepsilon) \rightarrow 0, \quad \text{for all } i = 1, ..., d.
\] (63)

Convergence in probability is invariant with respect to any continuous mapping:

Proposition 7 (Brockwell and Davis [6], Proposition 6.1.4). If \( \theta_n \overset{P}{\rightarrow} \theta_0 \) and \( g : \Omega \rightarrow \mathbb{R} \) is a continuous function at \( \theta = \theta_0 \), then \( g(\theta_n) \overset{P}{\rightarrow} g(\theta_0) \).

Definition 2. We say that a sequence \( \{ \theta_n \} \) converges in law (in distribution) to the random variable \( \tilde{\theta} \) and write \( \theta_n \overset{L}{\rightarrow} \tilde{\theta} \), iff the sequence of their joint CDFs, \( F_n \), point-wise converges to the joint CDF of \( \tilde{\theta} \):

\[
F_n(a) = P(\theta_{n1} \leq a_1, ..., \theta_{nd} \leq a_d) \rightarrow F(a) = P(\tilde{\theta}_1 \leq a_1, ..., \tilde{\theta}_d \leq a_d) \quad \forall a \in C_F \subseteq \mathbb{R}^d,
\] (64)

where \( C_F \) is the set of continuity points of the CDF \( F \).

Equation (64) means that for large values of \( n \), the distribution of \( \theta_n \) can be well approximated by the distribution of \( \tilde{\theta} \). Note that throughout this paper, for simplicity, we say that a random sequence \( \{ \theta_n \} \) converges to a distribution with density function \( p(\theta) \), or write \( \theta_n \overset{L}{\rightarrow} p(\theta) \), instead of fully saying that \( \{ \theta_n \} \) converges in law to a random variable with that distribution.

Note that \( \theta_n \overset{P}{\rightarrow} \theta_0 \) suggests that \( \theta_n - \theta_0 \overset{L}{\rightarrow} \delta(\theta) \) where \( \delta \) is the Kronecker delta function, which can be viewed as the density function of a degenerate distribution at \( \theta = 0 \). This, however,
does not give any information about the speed with which \( \theta_n \) converges to \( \theta_0 \). In order to take the speed into account, we consider the convergent distribution of the sequence \( \{a_n \cdot (\theta_n - \theta_0)\} \), where \( a_n \) is any sequence of positive integers and \( a_n \to \infty(n \to \infty) \). In practice \( a_n \) is usually considered to have the form \( n^r \) with \( r > 0 \).

**Definition 3.** Assume \( \theta_n \xrightarrow{P} \theta_0 \). We say that the sequence \( \{\theta_n\} \) converges to \( \theta_0 \) with rate of convergence \( r > 0 \), iff \( n^r(\theta_n - \theta_0) \) converges in law to a random variable with non-degenerate distribution. Furthermore, the non-degenerate distribution is the asymptotic distribution of \( \theta_n \).

Next, we discuss some of the classic results in asymptotic statistics:

**Theorem 8 (Law of Large Numbers, Brockwell and Davis [6]).** Let \( \theta_1, \ldots, \theta_n \) be a set of independent and identically distributed (i.i.d) samples. If \( \mathbb{E}[\theta_i] = \mu \), then

\[
\bar{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \theta_i \xrightarrow{P} \mu. 
\]  

(65)

**Theorem 9 (Central Limit Theorem, Lehmann and Casella [31]).** Let \( \theta_1, \ldots, \theta_n \) be a set of i.i.d samples with mean \( \mathbb{E}[\theta_i] = \mu \) and covariance \( \text{Cov}[\theta_i] = \Sigma \) (with a symmetric and positive semi-definite matrix \( \Sigma \)), then the sequence of sample averages \( \{\bar{\theta}_n\} \) with \( \theta_n = \frac{1}{n} \sum_{i=1}^{n} \theta_i \) converges to the true mean with convergence rate 1/2. Moreover, its asymptotic distribution is a zero-mean Gaussian distribution with covariance matrix \( \Sigma \), that is:

\[
\sqrt{n} \cdot (\bar{\theta}_n - \mu) \xrightarrow{L} \mathcal{N}(0, \Sigma).
\]  

(66)

The following results are very useful when deriving the asymptotic distribution of a random sequence under a continuous mapping:

**Theorem 10. (Multivariate Delta Method, first order, Lehmann and Casella [31])** Let \( \{\theta_n\} \) be a sequence of random variables such that it converges to \( \theta_0 \) with rate of convergence 1/2 and a normal asymptotic distribution, that is \( \sqrt{n} \cdot (\theta_n - \theta_0) \xrightarrow{L} \mathcal{N}(0, \Sigma) \). If \( g : \mathbb{R}^d \to \mathbb{R} \) is a continuously differentiable mapping and \( \nabla g(\theta_0) \neq 0 \), then

\[
\sqrt{n} \cdot \left[ g(\theta_n) - g(\theta_0) \right] \xrightarrow{L} \mathcal{N} \left( 0, \nabla^\top g(\theta_0) \Sigma \nabla g(\theta_0) \right).
\]  

(67)

**Theorem 11 (Multivariate Delta Method, second order).** Let \( \{\theta_n\} \) be a sequence of random variables such that it converges to \( \theta_0 \) with rate of convergence 1/2 and a normal asymptotic distribution, that is \( \sqrt{n} \cdot (\theta_n - \theta_0) \xrightarrow{L} \mathcal{N}(0, \Sigma) \). If \( g : \mathbb{R}^d \to \mathbb{R} \) is a continuously differentiable mapping where \( \nabla g(\theta_0) = 0 \) and \( \nabla^2 g(\theta_0) \) is non-singular in a neighborhood of \( \theta_0 \), then the sequence \( \{g(\theta_n) - g(\theta_0)\} \) converges in law to a mixture of random variables with first-degree Chi-square distributions, and the rate of convergence is one. More specifically,

\[
n \cdot \left[ g(\theta_n) - g(\theta_0) \right] \xrightarrow{L} \sum_{i,j=1}^{d} [\gamma_{ij} \chi_1^2],
\]  

(68)

where

\[
\gamma_{ij} = \left[ \Sigma^{1/2} \nabla g(\theta_0) \Sigma^{1/2} \right]_{ij}.
\]  

(69)
Moreover, variance of this asymptotic distribution can be written as

\[ \frac{1}{2} \left\| \Sigma^{1/2} \nabla_x^2 g(x_0) \Sigma^{1/2} \right\|_F^2, \tag{70} \]

where \( \left\| \cdot \right\|_F \) is the Frobenius norm.

Proof. For proof see Appendix B. \qed

A.2 Parameter Estimation

Now suppose that the set of independent and identically distributed (i.i.d) set of samples \( x_1, \ldots, x_n \) are generated from an underlying distribution that belongs to a parametric family, for which the density function \( p(x | \theta) \) can be represented by a multivariate parameter vector \( \theta \). Assume the true parameter is \( \theta_0 \), that is \( \{x_i\} \sim p(x | \theta_0), i = 1, \ldots, n \). An estimator \( \hat{\theta}_n = \theta(x_1, \ldots, x_n) \) is a function that maps the observed random variables to a point in the parameter space \( \Omega \). The subscript \( n \) in \( \hat{\theta}_n \) indicates its dependence on the sample size. Since the observations are generated randomly, the estimators are also random and thus \( \{\hat{\theta}_n\} \) can be viewed as a sequence of random variables. There are some reserved terms for such a sequence, which we introduce in the remaining of this section:

Definition 4 (Consistency). We say that an estimator \( \hat{\theta}_n \) is consistent iff \( \hat{\theta}_n \xrightarrow{P} \theta_0 \).

Based on Theorem \( \ref{consistency} \) sample average of the observation set is a consistent estimator of the true mean of the samples. Another important characteristic of estimators is based on the following bound over their covariance matrices:

Theorem 12 (Cramér-Rao, Lehmann and Casella). Let \( x_1, \ldots, x_n \sim p(x | \theta_0) \) and \( \hat{\theta}_n = \theta(x_1, \ldots, x_n) \) be an estimator. If the first moment of \( \hat{\theta}_n \) is differentiable with respect to the parameter vector and its second moment is finite, then the following inequality holds for every \( \theta \in \Omega \):

\[ \text{Cov}[\hat{\theta}_n] \succeq - (\nabla_\theta \mathbb{E}[\hat{\theta}_n])^\top \mathbb{I}(\theta)^{-1} \nabla_\theta \mathbb{E}[\hat{\theta}_n]. \tag{71} \]

The right-hand-side of (71) is called the Cramer-Rao bound of the estimator, where the middle term is the inverse of the Fisher information matrix of the parametric distribution \( p(x | \theta) \):

\[ \mathbb{I}(\theta) = \mathbb{E} \left[ \nabla_\theta \log p(x | \theta) \cdot \nabla_\theta^\top \log p(x | \theta) \right] \]

Theorem 12 suggests that for an unbiased estimator \( \hat{\theta}_n \), the inequality over the covariance matrix becomes: \( \text{Cov}[\hat{\theta}_n] \succeq \mathbb{I}(\theta)^{-1}, \forall \theta \in \Omega. \)

Definition 5 (Efficiency). We say that an estimator \( \hat{\theta}_n \) is efficient, iff it attains the Cramer-Rao bound, that is \( \text{Cov}[\hat{\theta}_n] \) achieves the lower-bound in (71) for every \( n = 1, 2, \ldots \). Furthermore, we say that \( \hat{\theta}_n \) is asymptotically efficient, iff the lower bound is attained asymptotically (when \( n \to \infty \)).

B Proof of Second-order Multivariate Delta Method

In order to prove this theorem, we have to formulate the statistical Taylor expansion. This, in turn, needs a brief introduction of stochastic order notations.
B.1 Stochastic Order Notations

The stochastic order notations are denoted by \( o_p \) and \( O_p \), where the former is equivalent to convergence in probability (Definition 1) and the latter implies boundedness in probability. In what follows, if otherwise stated, \( \{\theta_n\} \) is a sequence of multivariate random variables lying in \( \Omega \subseteq \mathbb{R}^d \) and \( \{a_n\} \) is a sequence of strictly positive real numbers. The skipped proofs can be found in many textbooks on asymptotic theory, such as Brockwell and Davis [6, Chapter 6].

**Definition 6.** We write \( \theta_n = o_p(a_n) \) iff

\[
\frac{\theta_{in}}{a_n} = o_p(1), \quad \text{for all } i = 1, \ldots, d
\]

(72)

**Definition 7.** We write \( \theta_n = O_p(a_n) \) iff the sequence \( \{\theta_{in}\} \) is bounded in probability for every \( i = 1, \ldots, d \), that is, for every \( \varepsilon > 0 \) there exists \( \delta_\varepsilon \) and \( N_\varepsilon \) such that

\[
P\left(\left|\frac{\theta_{in}}{a_n}\right| > \delta_\varepsilon\right) < \varepsilon, \quad \forall n \geq N_\varepsilon
\]

(73)

We also need the following propositions:

**Proposition 13** (Brockwell and Davis [6]). Let \( \{\theta_n\} \) and \( \{\eta_n\} \) be two sequences of scalar random variables, and \( \{a_n\} \) and \( \{b_n\} \) be two sequences of positive real numbers. If \( \theta_n = O_p(a_n) \) and \( \eta_n = o_p(b_n) \), then

(i) \( \theta_n^2 = O_p(a_n^2) \)

(ii) \( \theta_n \eta_n = o_p(a_n b_n) \)

**Proposition 14.** The followings are true\(^{10}\):

(i) \( \theta_n = o_p(a_n) \iff \|\theta_n\| = o_p(a_n) \).

(ii) \( \theta_n = O_p(a_n) \iff \|\theta_n\| = O_p(a_n) \).

**Proof.** The proof of part (i) can be found in Brockwell and Davis [6, Proposition 6.1.2]. Here, we only prove part (ii).

(ii, \( \Rightarrow \)) : Since \( \theta_n = O_p(a_n) \), for every \( \varepsilon > 0 \) and for every \( i = 1, \ldots, d \), there exists a coefficient \( \delta_i > 0 \) such that

\[
P(|\theta_{ni}| > a_n \cdot \delta_i) < \frac{\varepsilon}{d}, \quad n = 1, 2, \ldots
\]

(74)

Define \( \delta_{\max} = \max\{\delta_1, \ldots, \delta_d\} \) and note that we can write

\[
\left\{\theta_n : \sum_{i=1}^d |\theta_{ni}|^2 > (d \cdot a_n \cdot \delta_{\max})^2\right\} \subseteq \bigcap_{i=1}^d \left\{\theta_n : |\theta_{ni}| \leq a_n \cdot \delta_{\max}\right\}^c
\]

\[
= \bigcup_{i=1}^d \left\{\theta_n : |\theta_{ni}| > a_n \cdot \delta_{\max}\right\}
\]

(75)

\(^{10}\)Unless subscripted, \( \| \cdot \| \) denotes the \( L_2 \) norm in all the equations.
implying that
\[
P \left( \| \theta_n \|^2 > (d \cdot a_n \cdot \delta_{max})^2 \right) \leq P \left( \cup_{i=1}^{d} \{ \theta_n : |\theta_{ni}| > a_n \cdot \delta_{max} \} \right)
\]
\[
\leq \sum_{i=1}^{d} P(|\theta_{ni}| > a_n \cdot \delta_{max})
\]
(76)

Furthermore, for every \( i = 1, \ldots, d \) we have \( \delta_{max} \geq \delta_i \), consequently the interval \((a_n \delta_{max}, \infty)\) is a subset of \((a_n \delta_i, \infty)\) and \( P(|\theta_{ni}| > a_n \delta_{max}) \leq P(|\theta_{ni}| > a_n \delta_i) \). This implies that
\[
P \left( \| \theta_n \|^2 > (d \cdot a_n \cdot \delta_{max})^2 \right) \leq \sum_{i=1}^{d} P(|\theta_{ni}| > a_n \cdot \delta_i) < \varepsilon.
\]
(77)

Therefore, for every \( \varepsilon > 0 \), we can choose \( \delta_\varepsilon = d \cdot \delta_{max} \) such that \( P \left( \frac{\| \theta_n \|}{a_n} > \delta_\varepsilon \right) < \varepsilon \) for every \( n = 1, 2, \ldots \), that is \( \| \theta_n \| = O_p(a_n) \).

\( (ii, \Leftarrow) \): Suppose \( \| \theta_n \| = O_p(a_n) \), that is for every \( \varepsilon > 0 \) we can find \( \delta_\varepsilon > 0 \) such that
\[
P(\| \theta_n \| > a_n \cdot \delta_\varepsilon) < \varepsilon, \ n = 1, 2, \ldots
\]
(78)

It is clear that for any given \( i \in \{1, \ldots, d\} \) we have
\[
\{ \theta_n : |\theta_{ni}| > a_n \cdot \delta_\varepsilon \} \subseteq \{ \theta_n : \| \theta_n \| > a_n \cdot \delta_\varepsilon \}
\]
(79)

hence
\[
P(|\theta_{ni}| > a_n \cdot \delta_\varepsilon) \leq P(\| \theta_n \| > a_n \cdot \delta_\varepsilon) < \varepsilon, \ n = 1, 2, \ldots
\]
(80)

meaning that \( \theta_{ni} = O_p(a_n), i = 1, \ldots, d \) or equivalently \( \theta_n = O_p(a_n) \).

\[\Box\]

**Proposition 15.** If \( \theta_n = O_p(a_n) \) and \( a_n \to 0(n \to \infty) \), then \( \theta_n = o_p(1) \).

**Proof.** The goal is to show \( \theta_n = o_p(1) \) or equivalently \( \| \theta_n \| = o_p(1) \) by proving that \( P(\| \theta_n \| > \varepsilon) \to 0(n \to \infty) \) for every \( \varepsilon > 0 \). Fix \( \varepsilon \) to a positive real number. In order to have the sequence of probability numbers \( \{P(\| \theta_n \| > \varepsilon)\} \) converging to zero, for every \( \varepsilon_0 > 0 \) there should exist a positive integer \( N > 0 \) such that
\[
P(\| \theta_n \| > \varepsilon) < \varepsilon_0 \ \forall n > N.
\]
(81)

Because of the assumption of being bounded by \( a_n \), that is \( \theta_n = O_p(a_n) \) or equivalently \( \| \theta_n \| = O_p(a_n) \), we can choose a real number \( \delta_0 > 0 \) such that
\[
P(\| \theta_n \| > a_n \delta_0) < \varepsilon_0 \ \ n = 1, 2, \ldots
\]
(82)

On the other hand, since \( a_n \to 0(n \to \infty) \), there exists a large enough number \( N_0 > 0 \) such that \( 0 < a_n < \frac{\varepsilon}{\delta_0} \) for all \( n > N_0 \). Therefore we get:
\[
[0, a_n \delta_0] \subseteq [0, \varepsilon] \ \forall n > N_0
\]
(83)
implying that
\[ P(\| \theta_n \| \leq a_n \delta_0) \leq P(\| \theta_n \| \leq \varepsilon) \quad \forall n > N_0 \] (84)

From inequalities (82) and (84), and noticing that the latter holds for all \( n \) whereas the former is satisfied when \( n > N_0 \), one can write:
\[ P(\| \theta_n \| > \varepsilon) \leq P(\| \theta_n \| > a_n \delta_0) < \varepsilon_0 \quad \forall n > N_0 \] (85)

Therefore, for every \( \varepsilon_0 > 0 \), equation (81) is guaranteed if \( N \) is chosen to be equal to \( N_0 \) so that inequality (84) is satisfied. Similarly, this can be written for every \( \varepsilon > 0 \), thus the proof is complete.

**Proposition 16** (Serfling [39], Chapter 1). Let \( \{ \theta_n \} \) be a sequence of random variables. If there exists a random variable \( \theta \) such that \( \theta_n \xrightarrow{L} \theta \), then \( \theta_n = O_p(1) \).

**B.2 Second-order Statistical Taylor Expansion**

Now we are ready to establish the second-order statistical Taylor expansion.

**Theorem 17.** Let \( \{ \theta_n \} \) be a sequence of random vectors in a convex and compact set \( \Omega \subseteq \mathbb{R}^d \) and \( \theta_0 \in \Omega \) be a constant vector such that \( \theta_n - \theta_0 = O_p(a_n) \) where \( a_n \to 0(n \to \infty) \). If \( g : \Omega \to \mathbb{R} \) is a \( C^3 \) function, then
\[ g(\theta_n) = g(\theta_0) + (\theta_n - \theta_0)^	op \nabla g(\theta_0) + \frac{1}{2}(\theta_n - \theta_0)^	op \nabla^2 g(\theta_0)(\theta_n - \theta_0) + o_p(a_n^2). \] (86)

**Proof.** Since \( g \) is twice continuously differentiable in a neighborhood of \( \theta_0 \), it can be written in terms of the Taylor expansion as
\[ g(\theta) = g(\theta_0) + (\theta - \theta_0)^	op \nabla g(\theta_0) + \frac{1}{2}(\theta - \theta_0)^	op \nabla^2 g(\theta_0)(\theta - \theta_0) + r_2(\theta, \theta_0) \] (87)

where \( r_2(\theta, \theta_0) \) is the Lagrange remainder of second order. Based on Taylor’s polynomial theorem for multivariate functions, there exists a number \( t \in [0,1] \) such that \( \theta^* = t \theta + (1 - t) \theta_0 \in \Omega \) (due to convexity of \( \Omega \)) and
\[ r_2(\theta, \theta_0) = \sum_{1 \leq i,j,k \leq d} \frac{\partial^3 g(\theta^*)}{\partial \theta_i \partial \theta_j \partial \theta_k} (\theta_i - \theta_0i)(\theta_j - \theta_0j)(\theta_k - \theta_0k) \] (88)

But since \( \Omega \) is compact and \( g \in C^3 \), the third derivative of \( g \) is bounded\(^{11}\) and therefore there exists \( M > 0 \) such that
\[ \left| \frac{\partial^3 g(\theta)}{\partial \theta_i \partial \theta_j \partial \theta_k} \right| \leq M, \forall \theta \in \Omega, \forall i, j, k \in \{1, ..., d\} \] (89)

\(^{11}\)This is because of the following Theorem in real analysis:

**Theorem 18.** Let \( X \) and \( Y \) be two vector spaces. If \( g : X \to Y \) is continuous and \( X \) is compact, then \( f(X) \) is compact in \( Y \).

In special case of this theorem, when \( Y = \mathbb{R} \), compactness of \( f(X) \) is equivalent to boundedness and closedness.
Hence the Lagrange remainder can be bounded by
\[
|r_2(\theta, \theta_0)| \leq \frac{M}{6} \sum_{1 \leq i, j, k \leq 3} |\theta_i - \theta_0i| \cdot |\theta_j - \theta_0j| \cdot |\theta_k - \theta_0k|
\]
\[
= \frac{M}{6} \|\theta - \theta_0\|_1^3
\]
\[
\leq \frac{M'}{6} \|\theta - \theta_0\|_2^3
\]
(90)
where \(M' = c_uM\) and \(c_u\) is obtained from the equivalence of norms in \(\mathbb{R}^d\). Now define the function \(h : \Omega \to \mathbb{R}\) as below
\[
h(\theta) := \begin{cases} r_2(\theta, \theta_0) / \|\theta - \theta_0\|_2^2, & \theta \neq \theta_0 \\ 0, & \theta = \theta_0 \end{cases}
\]
(92)
Note that \(h(\theta)\) is continuous at \(\theta = \theta_0\): due to boundedness of \(r_2(\theta, \theta_0)\), \(h(\theta)\) is also bounded by
\[
|h(\theta)| \leq \frac{M'}{3} \|\theta - \theta_0\|.
\]
(93)
Hence, for every \(\varepsilon > 0\), we can select \(\delta = \frac{3 \varepsilon}{M'}\) such that the following continuity condition holds
\[
\|\theta - \theta_0\| < \delta \Rightarrow |h(\theta)| \leq \varepsilon.
\]
(94)
Continuity of \(h(\theta)\) at \(\theta = \theta_0\) implies \(\lim_{\theta \to \theta_0} h(\theta) = h(\theta_0) = 0\). Furthermore, since \(\theta_n - \theta_0 = O_p(a_n)\) and \(a_n \to 0(n \to \infty)\), Proposition 15 suggests that \(\theta_n - \theta_0 = o_p(1)\). These two enable us to use Proposition 7 and write
\[
h(\theta_n) - h(\theta_0) = h(\theta_n) = o_p(1).
\]
(95)
Finally, from equation (92) and Propositions 13, 14 and 15, we can write that
\[
r_2(\theta_n, \theta_0) = h(\theta_n) \cdot \frac{\|\theta_n - \theta_0\|^2}{2} = o_p(1) \cdot O_p(a_n^2) = o_p(a_n^2)
\]
(96)

**B.3 Second-order Multivariate Delta Method**

Finally, here is the proof of second-order multivariate Delta method (Theorem 11):

**Proof.** From assumption of the Theorem, \(\sqrt{n}(\theta_n - \theta_0) \xrightarrow{L} \mathcal{N}(0, \Sigma)\), and Proposition 16 one conclude that \(\sqrt{n}(\theta_n - \theta_0) = O_p(1)\) and therefore \(\theta_n - \theta_0 = O_p\left(\frac{1}{\sqrt{n}}\right)\). Thus we can use Theorem 17 with \(a_n = \frac{1}{\sqrt{n}}\) to write:
\[
g(\theta_n) = g(\theta_0) + (\theta_n - \theta_0)^\top \nabla g(\theta_0) + \frac{1}{2}(\theta_n - \theta_0)^\top \nabla^2 g(\theta_0)(\theta_n - \theta_0) + o_p\left(\frac{1}{\sqrt{n}}\right),
\]
\[
\]
\[
n \left[g(\theta_n) - g(\theta_0)\right] = \frac{1}{2} \left[\sqrt{n} \cdot (\theta_n - \theta_0)\right]^\top \nabla^2 g(\theta_0) \left[\sqrt{n} \cdot (\theta_n - \theta_0)\right] + o_p(1)
\]
\[
\overset{\text{L}}{\sim} \frac{1}{2} \mathcal{N}(0, \Sigma)^\top \nabla^2 g(\theta_0) \mathcal{N}(0, \Sigma)
\]
\[
= \frac{1}{2} \mathcal{N}(0, I_d)^\top \left[\Sigma^{1/2} \nabla^2 g(\theta_0) \Sigma^{1/2}\right] \mathcal{N}(0, I_d)
\] (98)

Define \( \Gamma := \Sigma^{1/2} \nabla^2 g(\theta_0) \Sigma^{1/2} \) and rewrite the right-hand-side element-wise as
\[
\frac{1}{2} \mathcal{N}(0, I_d)^\top \Gamma \mathcal{N}(0, I_d) = \frac{1}{2} \sum_{i,j=1}^d \gamma_{ij} \mathcal{N}(0, 1)^2 = \frac{1}{2} \sum_{i,j=1}^d \gamma_{ij} \chi_1^2,
\] (99)

where
\[
\gamma_{ij} = [\Gamma]_{ij} = \left[\Sigma^{1/2} \nabla^2 g(\theta_0) \Sigma^{1/2}\right]_{ij}.
\]

Finally, noting that the terms in the Chi-square mixture are independent, variance of the convergent random variable can be easily computed as
\[
\text{Var} \left[\frac{1}{2} \sum_{i,j=1}^d \gamma_{ij} \mathcal{N}(0, 1)^2\right] = \frac{1}{4} \sum_{i,j=1}^d \gamma_{ij}^2 \cdot \text{Var} [\chi_1^2]
\]
\[
= \frac{1}{2} \sum_{i,j=1}^d \gamma_{ij}^2
\]
\[
= \frac{1}{2} \left\|\Sigma^{1/2} \nabla^2 g(x_0) \Sigma^{1/2}\right\|_F^2,
\] (100)

\square

References

[1] P. Awasthi, M. F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. arXiv preprint arXiv:1307.8371, 2013.

[2] J. Azimi, A. Fern, X. Zhang-Fern, G. Borrodaile, and B. Heeringa. Batch active learning via coordinated matching. arXiv preprint arXiv:1206.6458, 2012.

[3] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In Proceedings of the 23rd International Conference on Machine Learning, pages 65–72. ACM, 2006.

[4] A. Beygelzimer, J. Langford, Z. Tong, and D. J. Hsu. Agnostic active learning without constraints. In Advances in Neural Information Processing Systems, pages 199–207, 2010.

[5] C. M. Bishop et al. Pattern Recognition and Machine Learning, volume 1. Springer New York, 2006.
[6] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer, 1991.

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

[8] Y. Chen and A. Krause. Near-optimal batch mode active learning and adaptive submodular optimization. In *Proceedings of The 30th International Conference on Machine Learning*, pages 160–168, 2013.

[9] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

[10] D. A. Cohn. Neural network exploration using optimal experiment design. *Neural Networks*, 9(6):1071–1083, 1996.

[11] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *arXiv preprint cs/9603104*, 1996.

[12] S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems*, pages 235–242, 2005.

[13] S. Dasgupta, A. T. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *Learning Theory*, pages 249–263. Springer, 2005.

[14] S. Dasgupta, C. Monteleoni, and D. J. Hsu. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems*, pages 353–360, 2007.

[15] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons,, 1999.

[16] V. V. Fedorov. *Theory of Optimal Experiments*. Elsevier, 1972.

[17] A. Feiveson and F. Delaney. *The distribution and properties of a weighted sum of chi squares*. National Aeronautics and Space Administration; for sale by the Clearinghouse for Federal Scientific and Technical Information, Springfield, VA., 1968.

[18] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[19] Y. Fu, X. Zhu, and B. Li. A survey on instance selection for active learning. *Knowledge and information systems*, 35(2):249–283, 2013.

[20] K. Fukumizu. Statistical active learning in multilayer perceptrons. *IEEE Transactions on Neural Networks*, 11(1):17–26, 2000.

[21] Q. Gu, T. Zhang, and J. Han. Batch-mode active learning via error bound minimization. *Urbana*, 51:61801, 2014.

[22] Y. Guo. Active instance sampling via matrix partition. In *Advances in Neural Information Processing Systems*, pages 802–810, 2010.
[23] Y. Guo and R. Greiner. Optimistic active-learning using mutual information. In National Joint Conferences on Artificial Intelligence, volume 7, pages 823–829, 2007.

[24] Y. Guo and D. Schuurmans. Discriminative batch mode active learning. In Advances in Neural Information Processing Systems, pages 593–600, 2008.

[25] S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. The Journal of Machine Learning Research, 98888(1):1469–1587, 2012.

[26] S. Hanneke et al. Rates of convergence in active learning. The Annals of Statistics, 39(1):333–361, 2011.

[27] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Batch mode active learning and its application to medical image classification. In Proceedings of the 23rd International Conference on Machine Learning, pages 417–424. ACM, 2006.

[28] S. C. Hoi, R. Jin, and M. R. Lyu. Batch mode active learning with applications to text categorization and image retrieval. IEEE Transactions on Knowledge and Data Engineering, 21(9):1233–1248, 2009.

[29] A. Holub, P. Perona, and M. C. Burl. Entropy-based active learning for object recognition. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Workshop on Online Learning for Classification, pages 1–8, 2008.

[30] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. The Journal of Machine Learning Research, 9:235–284, 2008.

[31] E. L. Lehmann and G. Casella. Theory of Point Estimation. Springer, 1998.

[32] D. MacKay. Information-based objective functions for active data selection. Neural Computation, 4(4):590–604, 1992.

[33] N. Murata, S. Yoshizawa, and S.-I. Amari. Network information criterion-determining the number of hidden units for an artificial neural network model. IEEE Transactions on Neural Networks, 5(6):865–872, 1994.

[34] G. L. Nemhauser and L. A. Wolsey. Best algorithms for approximating the maximum of a submodular set function. Mathematics of Operations Research, 3(3):177–188, 1978.

[35] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. Mathematical Programming, 14(1):265–294, 1978.

[36] H. T. Nguyen and A. Smeulders. Active learning using pre-clustering. In Proceedings of the 21st International Conference on Machine Learning, page 79. ACM, 2004.

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

[38] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In International Conference on Machine Learning, pages 839–846, 2000.
[39] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 2002.

[40] B. Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012. doi: 10.2200/S00429ED1V01Y201207AIM018.

[41] B. Settles and M. Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, 2008.

[42] B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. In *Advances in Neural Information Processing Systems*, pages 1289–1296, 2008.

[43] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[44] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[45] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.

[46] X. Yang. A matrix trace inequality. *Journal of Mathematical Analysis and Applications*, 250(1):372–374, 2000.

[47] C. Zhang and K. Chaudhuri. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems*, pages 442–450, 2014.

[48] T. Zhang and F. Oles. The value of unlabeled data for classification problems. In *Proceedings of the 17th International Conference on Machine Learning*, pages 1191–1198, 2000.

[49] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *International Conference on Machine Learning, Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, pages 58–65, 2003.