Journal of Machine Learning Research 18 (2017) 1-41 Submitted 3/15; Revised 2/17; Published 4/17

# Asymptotic Analysis of Objectives Based on Fisher Information in Active Learning

**Jamshid Sourati** SOURATI@ECE.NEU.EDU
*Department of Electrical and Computer Engineering*
*Northeastern University*
*Boston, MA 02115 USA*

**Murat Akcakaya** AKCAKAYA@PITT.EDU
*Department of Electrical and Computer Engineering*
*University of Pittsburgh*
*Pittsburgh, PA 15261 USA*

**Todd K. Leen** TODD.LEEN@GEORGETOWN.EDU
*Georgetown University*
*Washington D.C. 20057 USA*

**Deniz Erdogmus** ERDOGMUS@ECE.NEU.EDU
**Jennifer G. Dy** JDY@ECE.NEU.EDU
*Department of Electrical and Computer Engineering*
*Northeastern University*
*Boston, MA 02115 USA*

**Editor:** Qiang Liu

## Abstract

Obtaining labels can be costly and time-consuming. Active learning allows a learning algorithm to intelligently query samples to be labeled for a more efficient learning. Fisher information ratio (FIR) has been used as an objective for selecting queries. 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 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 allows us to give insight into the development of new active learning approaches based on this objective.

**Keywords:** classification active learning, Fisher information ratio, asymptotic log-loss, upper-bound minimization.

## 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 (Cohn et al., 1994). 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 efficiently trained learner by querying a fixed amount of knowledge (Settles, 2012; Fu et al., 2013).

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.*

From the statistical point of view, 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 (Duda et al., 1999), 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 (Bishop, 2006). In order to better categorize the existing techniques, we break an active learning algorithm into the following sub-problems:

(i) *(Query Selection)* Sampling a set of covariates $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ from the *training* marginal[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 subproblem is formulated in terms of the distribution from which the queries will be drawn.

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.

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 (Settles et al., 2008; Guo and Schuurmans, 2008) or Fisher information criterion (Hoi et al., 2006; Settles and Craven, 2008; Hoi et al., 2009; Chaudhuri et al., 2015b). On the other hand, the wide range of studies in prediction-based active learning includes a more varied set of objectives: for instance, the prediction error probability[2] (Cohn et al., 1994; Freund et al., 1997; Zhu et al., 2003; Nguyen and Smeulders, 2004; Dasgupta, 2005; Dasgupta et al., 2005; Balcan et al., 2006; Dasgupta et al., 2007; Beygelzimer et al., 2010; Hanneke et al., 2011; Hanneke, 2012; Awasthi et al., 2013; Zhang and Chaudhuri, 2014), variance of the predictions (Cohn et al., 1996; Schein and Ungar, 2007; Ji and Han, 2012), uncertainty of the learner with respect to the unknown labels as evaluated by the entropy function (Holub et al., 2008), mutual information (Guo and Greiner, 2007; Krause et al., 2008; Guo, 2010; Sourati et al., 2016), and margin of the samples with respect to the trained hyperplanar discriminant function (Schohn and Cohn, 2000; Tong and Koller, 2002).

In this paper, 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)* (Settles and Craven, 2008) and its usage is motivated by older attempts in optimal experiment design for statistical regression methods (Fedorov, 1972; MacKay, 1992; Cohn, 1996; Fukumizu, 2000).

Among the existing FIR-based classification querying methods, only the very first one proposed by Zhang and Oles (2000) 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 i above). Later on, Hoi et al. (2006) and Hoi et al. (2009), inspired by Zhang and Oles (2000), 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 (2008) employed this objective with the same motivation, but using a different approximation and optimization technique. More recently, Chaudhuri et al. (2015b) showed that even finite-sample FIR is closely related to the expected log-likelihood ratio of an MLE-based classifier. However, their results are derived under a different and rather restricting set of conditions and assumptions: they focused on the finite-sample case where the test marginal is a uniform PMF and the proposal marginal is a general PMF (to be determined) over a finite pool of unlabeled samples. Moreover, they assumed that the conditional Fisher information matrix is assumed to be independent of the class labels. Here, in a framework similar to Zhang and Oles (2000) but with a more

2. Prediction error probability is indeed the frequentist risk function of 0/1 loss, and is also known as *generalization error*.

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, and with no restricting assumptions on the form of test or training marginals, we show that FIR can be viewed as an upper bound for the expected variance of the asymptotic distribution of the log-likelihood ratio. Inspired by Chaudhuri et al. (2015b), we also show that under certain extra conditions, this relationship holds even in finite-sample case.

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 as in Fukumizu, 2000). 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 (Settles and Craven, 2008) and relaxation to continuous domains (Guo, 2010). Another strategy is to take advantage of *monotonic submodularity* of the objective set functions. If the objective is shown to be monotonically submodular, efficient greedy algorithms can be used for optimization with guaranteed tight bounds (Krause et al., 2008; Azimi et al., 2012; Chen and Krause, 2013). Regarding FIR, Hoi et al. (2006) 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 *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, implying 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 4.2);
- Proving that under certain assumptions, the above-mentioned asymptotic relationship also holds for finite-sample estimation of FIR (Section 5.1.1);
- Discussing different existing methods for coping with practical issues in using FIR in querying algorithms (Section 5.1), and accordingly providing a unifying framework for existing FIR-based active learning methods (Section 5.2).

- Proving submodularity for the Monte-Carlo simulation of FIR under *any* discriminative classifier, assuming a pool-based active learning which enables access to approximations of Fisher information matrices of both test and training distributions (Lemma 7 and Theorem 8).

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 $\{1, ..., c\}$. Suppose that the pairs $(\mathbf{x}, y)$ are distributed according to a parametric joint distribution $p(\mathbf{x}, y|\,\boldsymbol{\theta})$, with the parameter space denoted by $\Omega \subseteq \mathbb{R}^d$. Using a set of observed pairs as the training data, $\mathcal{L}_n := \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$, we can estimate $\boldsymbol{\theta}$ and predict the class labels of the unseen test samples, e.g., by maximizing $p(y|\,\mathbf{x}, \boldsymbol{\theta})$. In active learning, the algorithm is permitted to take part in designing $\mathcal{L}_n$ by choosing a set of data points $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$, 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 $\boldsymbol{\theta}$ comes only from the class-conditional distribution and the marginal distribution does not depend on $\boldsymbol{\theta}$, that is,

$$p(\mathbf{x}, y|\,\boldsymbol{\theta}) \;=\; p(y|\,\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}). \tag{1}$$

Zhang and Oles (2000) 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 $p(y|\,\mathbf{x}, \boldsymbol{\theta})$ (inference step in sub-problem ii).

(A1). *(Identifiability):* The joint distribution $P_{\boldsymbol{\theta}}$, whose density is given by $p(\mathbf{x}, y|\,\boldsymbol{\theta})$, is identifiable for different parameters. Meaning that for every distinct parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in $\Omega$, $P_{\boldsymbol{\theta}_1}$ and $P_{\boldsymbol{\theta}_2}$ are also distinct. That is,

$$\forall\, \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \in \Omega \;\; \exists A \subseteq X \times \{1, ..., c\} \quad \text{s.t.} \quad P_{\boldsymbol{\theta}_1}(A) \neq P_{\boldsymbol{\theta}_2}(A).$$

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

(A3). *(Model Faithfulness):* For any $\mathbf{x} \in X$, we have access to an oracle that generates a label $y$ according to the conditional $p(y|\,\mathbf{x}, \boldsymbol{\theta}_0)$. That is, the posterior parametric model matches the oracle distribution. We call $\boldsymbol{\theta}_0$ the true model parameter.

(A4). *(Training joint):* The set of observations in $\mathcal{L}_n := \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ are drawn independently from the *training/proposal/query* joint distribution of the form $p(y|\,\mathbf{x}, \boldsymbol{\theta}_0)$ $q(\mathbf{x})$, 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 $p(y|\,\mathbf{x}, \boldsymbol{\theta}_0)p(\mathbf{x})$ where $p$ is the test marginal with no dependence on the parameter.

(A6). *(Differentiability):* The log-conditional $\log p(y|\,\mathbf{x}, \boldsymbol{\theta})$ is of class $\mathcal{C}^3(\Omega)$ for all $(\mathbf{x}, y) \in X \times \{1, ..., c\}$, when being viewed as a funtion of the parameter.[3]

(A7). The parameter space $\Omega$ is compact and there exists an open ball around the true parameter of the model $\boldsymbol{\theta}_0 \in \Omega$.

(A8). *(Invertibility):* The Fisher information matrix (reviewed in Section 3.2) of the joint distribution is positive definite and therefore invertible for all $\boldsymbol{\theta} \in \Omega$, 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 necessarily 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 (2000) called *covariate shift in distribution*. In the remaining sections of the report, we use subscripts $p$ and $q$ for the statistical operators that consider $p(\mathbf{x})$ and $q(\mathbf{x})$ as the marginal in the joint distribution, respectively. We explicitly mention $\mathbf{x}$ as the input argument in order to refer to marginal operators. For instance, $\mathbb{E}_q$ denotes the joint expectation with respect to $q(\mathbf{x})p(y|\,\boldsymbol{\theta}, \mathbf{x})$, whereas $\mathbb{E}_{q(\mathbf{x})}$ denotes the marginal expectation with respect to $q(\mathbf{x})$.

## 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 subproblem (ii) with the assumptions listed above.

### 3.1 Maximum Likelihood Estimation

Here, we review maximum likelihood estimation in the context of classification problem. Given a training data set $\mathcal{L}_n = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$, a maximum likelihood estimate (MLE) is obtained by maximizing the log-likelihood function over all pairs inside $\mathcal{L}_n$, with respect to the parameter $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}}_n \;=\; \arg\max_{\boldsymbol{\theta}} \;\; \log p\left(\mathcal{L}_n \,|\, \boldsymbol{\theta}\right). \tag{2}$$

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

Under the assumptions (A0) and (A4), the optimization in (2) can be written as

$$\hat{\boldsymbol{\theta}}_n \;=\; \arg\max_{\boldsymbol{\theta}} \; \sum_{i=1}^{n} \log p(y_i|\, \mathbf{x}_i, \boldsymbol{\theta}). \tag{3}$$

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 $\boldsymbol{\theta}$, for any given pair $(\mathbf{x}, y) \in X \times \{1, ..., c\}$:

$$\ell(\boldsymbol{\theta}; \mathbf{x}, y) \;:=\; \log p(y|\, \mathbf{x}, \boldsymbol{\theta}). \tag{4}$$

Moreover, for any set of pairs independently generated from the joint distribution of the training data, such as $\mathcal{L}_n$ mentioned in (A4), the log-likelihood function will be

$$\ell(\boldsymbol{\theta}; \mathcal{L}_n) \;=\; \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \mathbf{x}_i, y_i) \;=\; \sum_{i=1}^{n} \log p(y_i|\, \mathbf{x}_i, \boldsymbol{\theta}). \tag{5}$$

Hence, the MLE can be rewritten as

$$\hat{\boldsymbol{\theta}}_n \;=\; \arg\max_{\boldsymbol{\theta}} \; \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \mathbf{x}_i, y_i). \tag{6}$$

Doing this maximization usually involves the computation of the stationary points of the log-likelihood, which requires calculating $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathcal{L}_n) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}_i, y_i)$. 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:

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}, y) \;=\; \nabla_{\boldsymbol{\theta}} \log p(y|\, \mathbf{x}, \boldsymbol{\theta}) \;=\; \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}, y|\, \boldsymbol{\theta}). \tag{7}$$

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.[4]

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

$$\hat{y}(\mathbf{x}) \;=\; \arg\max_{y} \ell(\hat{\boldsymbol{\theta}}_n; \mathbf{x}, y). \tag{8}$$

### 3.2 Fisher Information

In this section, 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 (1998).

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

4. Score function is actually zero-mean even under weaker regularity conditions.

of its importance to the Cramér-Rao Theorem (see Appendix A.2, Theorem 19), which guarantees a lower-bound for the covariance of the parameter estimators.

Fisher information, denoted by $\mathbf{I}(\boldsymbol{\theta})$, is defined as the expected value of the outer-product of the score function with itself, evaluated at some $\boldsymbol{\theta} \in \Omega$. 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:

$$\begin{aligned} \mathbf{I}_q(\boldsymbol{\theta}) &:= \mathbb{E}_q \left[ \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}, y|\, \boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}}^\top \log p(\mathbf{x}, y|\, \boldsymbol{\theta}) \right], \\ \mathbf{I}_p(\boldsymbol{\theta}) &:= \mathbb{E}_p \left[ \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}, y|\, \boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}}^\top \log p(\mathbf{x}, y|\, \boldsymbol{\theta}) \right]. \end{aligned} \tag{9}$$

Here, we focus on $\mathbf{I}_q$ to further explain Fisher information criterion. Our descriptions can be directly generalized to $\mathbf{I}_p$ as well. First, note that from equation (7) and that the score function is always zero-mean, one can reformulate the definition as

$$\begin{aligned} \mathbf{I}_q(\boldsymbol{\theta}) &= \mathbb{E}_q \left[ \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}, y) \cdot \nabla_{\boldsymbol{\theta}}^\top \ell(\boldsymbol{\theta}; \mathbf{x}, y) \right] \\ &= \mathrm{Cov}_q \left[ \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}, y) \right]. \end{aligned} \tag{10}$$

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:

$$\mathbf{I}_q(\boldsymbol{\theta}) = -\mathbb{E}_q \left[ \nabla_{\boldsymbol{\theta}}^2 \ell(\boldsymbol{\theta}; \mathbf{x}, y) \right]. \tag{11}$$

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

$$\begin{aligned} \mathbf{I}_q(\boldsymbol{\theta}) &= -\mathbb{E}_{q(\mathbf{x})} \left[ \mathbb{E}_{y|\,\mathbf{x},\boldsymbol{\theta}} \left[ \nabla_{\boldsymbol{\theta}}^2 \ell(\boldsymbol{\theta}; \mathbf{x}, y) |\, \mathbf{x}, \boldsymbol{\theta} \right] \right] \\ &= -\int_{\mathbf{x} \in X} q(\mathbf{x}) \left[ \sum_{y=1}^{c} p(y|\, \mathbf{x}, \boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}}^2 \ell(\boldsymbol{\theta}; \mathbf{x}, y) \right] d\,\mathbf{x}. \end{aligned} \tag{12}$$

### 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 the next section. More detailed descriptions of these properties, together with the proofs that are skipped here, can be found in different sources, such as Wasserman (2004) and Lehmann and Casella (1998).

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 (1998), Theorem 5.1)** *If the assumptions (A0) to (A7) hold, then there exists a sequence of solutions* $\left\{ \hat{\boldsymbol{\theta}}_n^* \right\}_{n=1}^{\infty}$ *to* $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathcal{L}_n) = 0$ *that converges to the true parameter* $\boldsymbol{\theta}_0$ *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 $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathcal{L}_n) = 0$ (the equation to solve for finding the stationary points) for every $n$, it is not obvious which root to select as $\hat{\boldsymbol{\theta}}_n^*$ to sustain the convergence. Therefore, while consistency of the MLE is guaranteed for models with a unique root of the score function evaluated at $\mathcal{L}_n$, 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, or become asymptotically unique, or we have access to an external procedure guiding us to select the proper roots so that $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$. We will denote the selected roots the same as $\hat{\boldsymbol{\theta}}_n$ from now on.

**Theorem 2 (Lehmann and Casella 1998, Theorem 5.1)** *Let $\hat{\boldsymbol{\theta}}_n$ be the maximum likelihood estimator based on the training data set $\mathcal{L}_n$. If the assumptions (A0) to (A8) hold, then the MLE $\hat{\boldsymbol{\theta}}_n$ has a zero-mean normal asymptotic distribution with the covariance equal to the inverse Fisher information matrix, and with the convergence rate of $1/2$:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{L} \mathcal{N}\left(\mathbf{0}, \mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}\right). \tag{13}$$

Theorems 2 and Cramér-Rao bound (see Appendix A), together with the consistency assumption, i.e., $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$, imply that MLE is an asymptotically efficient estimator with the efficiency equal to the training Fisher information. One can rewrite (13) as

$$\sqrt{n} \cdot \mathbf{I}_q(\boldsymbol{\theta}_0)^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \mathbb{I}_d). \tag{14}$$

In the following corollary, we see that if we substitute $\mathbf{I}_q(\boldsymbol{\theta}_0)$ with $\mathbf{I}_q(\hat{\boldsymbol{\theta}}_n)$, the new sequence still converges to a normal distribution:

**Corollary 3 (Wasserman 2004, Theorem 9.18)** *Under the assumptions given in Theorem 2, we get*

$$\sqrt{n} \cdot \mathbf{I}_q(\hat{\boldsymbol{\theta}}_n)^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{L} \mathcal{N}\left(\mathbf{0}, \mathbb{I}_d\right). \tag{15}$$

## 4. Fisher Information Ratio as an Upper Bound

In this section, we give our main theoretical analysis to relate FIR to the asymptotic distribution of the parameter log-likelihood ratio. Using the established relationship, we then show that FIR can be viewed as an asymptotic upper-bound of the expected variance of the loss function.

### 4.1 Asymptotic Distribution of MLE-Based Classifier

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

$$\ell(\hat{\boldsymbol{\theta}}_n; \mathbf{x}, y) \;-\; \ell(\boldsymbol{\theta}_0; \mathbf{x}, y). \tag{16}$$

This ratio can be viewed as an example of the classification loss function whose expectation with respect to the test joint distribution of $\mathbf{x}$ and $y$, results in the *discrepancy* between the

true conditional $p(y|\,\mathbf{x},\boldsymbol{\theta}_0)$ and MLE conditional $p(y|\,\mathbf{x},\hat{\boldsymbol{\theta}}_n)$ (Murata et al., 1994). 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 $(\mathbf{x}, y)$.

Furthermore, equation (16) is dependent on both the true marginal $p(\mathbf{x})$ (through the test pairs, where it should be evaluated) and the proposal marginal $q(\mathbf{x})$ (through the MLE $\hat{\boldsymbol{\theta}}_n$). In the classification context, Zhang and Oles (2000) claimed that the expected value of this ratio with respect to both marginals converges to $\mathrm{tr}[\mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}\,\mathbf{I}_p(\boldsymbol{\theta}_0)]$ with the convergence rate equal to unity. In the scalar case, $\mathrm{tr}[\mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}\,\mathbf{I}_p(\boldsymbol{\theta}_0)]$ 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* (Settles and Craven, 2008). This objective has been widely studied in linear and non-linear regression problems (Fedorov, 1972; MacKay, 1992; Murata et al., 1994; Cohn, 1996; Fukumizu, 2000). However, it is not as fully analyzed in classification.

Zhang and Oles (2000) and many papers following them (Hoi et al., 2006; Settles and Craven, 2008; Hoi et al., 2009), 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 4** *If the assumptions (A0) to (A8) hold, then, at any given* $(\mathbf{x}, y) \in X \times \{1, ..., c\}$*:*

*(I) In case* $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_0; \mathbf{x}, y) \neq \mathbf{0}$*, the log-likelihood ratio follows an asymptotic normality with convergence rate equal to* $1/2$*. More specifically,*

$$\sqrt{n}\cdot\Big(\ell(\hat{\boldsymbol{\theta}}_n; \mathbf{x}, y) - \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)\Big) \;\xrightarrow{L}\; \mathcal{N}\Big(0, \mathrm{tr}\big[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_0; \mathbf{x}, y)\cdot\nabla_{\boldsymbol{\theta}}^{\top}\ell(\boldsymbol{\theta}_0; \mathbf{x}, y)\cdot\mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}\big]\Big). \tag{17}$$

*(II) In case* $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_0; \mathbf{x}, y) = \mathbf{0}$ *and* $\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_0; \mathbf{x}, y)$ *is non-singular, the asymptotic distribution of the log-likelihood ratio is a mixture of Chi-square distributions with one degree of freedom, and the convergence rate is one. More specifically,*

$$n\cdot\Big(\ell(\hat{\boldsymbol{\theta}}_n; \mathbf{x}, y) \;-\; \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)\Big) \;\xrightarrow{L}\; \sum_{i=1}^{d}\lambda_i\cdot\chi^2_1, \tag{18}$$

*where* $\lambda_i$*'s are eigenvalues of* $\mathbf{I}_q(\boldsymbol{\theta}_0)^{-1/2}\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_0; \mathbf{x}, y)\,\mathbf{I}_q(\boldsymbol{\theta}_0)^{-1/2}$*.*

**Proof** Due to assumptions (A0) to (A7), Theorem 2 holds and therefore we have $\sqrt{n}\cdot(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \mathbf{I}_q(\boldsymbol{\theta}_0)^{-1})$. The rest of the proof is based on the Delta method in the two modes described in Appendix A (Theorems 16 and 17):

(I) $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_0; \mathbf{x}, y) \neq \mathbf{0}$ :

Since the expected log-likelihood function, evaluated at a given pair $(\mathbf{x}, y)$, is assumed to be continuously differentiable (A6) and that $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) \neq \mathbf{0}$, we can apply Theorem 16 to $\ell(\hat{\boldsymbol{\theta}}_n; \mathbf{x}, y) - \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)$ to write:

$$\sqrt{n} \cdot \Big( \ell(\hat{\boldsymbol{\theta}}_n; \mathbf{x}, y) \;-\; \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) \Big) \;\xrightarrow{L}\; \mathcal{N}\left( 0 \;,\; \nabla_{\boldsymbol{\theta}}^{\top} \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) \cdot \mathbf{I}_q(\boldsymbol{\theta}_0)^{-1} \cdot \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) \right), \tag{19}$$

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

(II) $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) = \mathbf{0}$ and $\nabla^2_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)$ non-singular :

In this case, the conditions in Theorem 17 are satisfied (with $\boldsymbol{\Sigma} = \mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}$ and $g(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}; \mathbf{x}, y)$), and therefore we can directly write (18) from equations (51).

■

Theorem 4 regards the log-likelihood ratio (16) evaluated at any arbitrary pair $(\mathbf{x}, y)$. Note that if we consider the training pairs in $\mathcal{L}_n$, which are used to obtain $\hat{\boldsymbol{\theta}}_n$, it is known that the ratio evaluated at the training set converges to a single Chi-square distribution with degree one, that is,

$$\ell(\hat{\boldsymbol{\theta}}_n; \mathcal{L}_n) \;-\; \ell(\boldsymbol{\theta}_0; \mathcal{L}_n) \;\xrightarrow{L}\; \frac{1}{2}\chi^2_1. \tag{20}$$

Theorem 4 implies that variance of the asymptotic distribution of the log-likelihood ratio in case (I) is $\operatorname{tr}\big[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) \cdot \nabla_{\boldsymbol{\theta}}^{\top} \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) \cdot \mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}\big]$, whereas in case (II), from Theorem 17 (see Appendix A), the variance is $\frac{1}{2} \left\| \mathbf{I}_q(\boldsymbol{\theta}_0)^{-1/2} \nabla^2_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)\, \mathbf{I}_q(\boldsymbol{\theta}_0)^{-1/2} \right\|_F^2$. Therefore, it is evident that the variance of the log-likelihood ratio at any $(\mathbf{x}, y)$ 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.

### 4.2 Establishing the Upper Bound

In the next theorem, we show that the Fisher information ratio , $\operatorname{tr}\left[\mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}\, \mathbf{I}_p(\boldsymbol{\theta}_0)\right]$, is a reasonable candidate objective to minimize in order to get a training distribution $q$ for the multivariate case.

**Theorem 5** *If the assumptions (A0) to (A8) hold, then*

$$\mathbb{E}_p \left[ \operatorname{Var}_q \left( \lim_{n \to \infty} \sqrt{n} \cdot [\ell(\hat{\boldsymbol{\theta}}_n; \mathbf{x}, y) - \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)] \right) \right] \;\leq\; \operatorname{tr}\left[ \mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}\, \mathbf{I}_p(\boldsymbol{\theta}_0) \right]. \tag{21}$$

*The equality holds if the set of pairs $(\mathbf{x}, y)$ where we have zero score function at $\boldsymbol{\theta}_0$, i.e., $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) = \mathbf{0}$, has measure zero under the true joint distribution $P_{\boldsymbol{\theta}_0}$ in $X \times \{1, ..., c\}$.*

**Proof** First note that $\operatorname{Var}_q \left( \lim_{n\to\infty} \sqrt{n} \cdot [\ell(\hat{\boldsymbol{\theta}}_n; \mathbf{x}, y) - \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)] \right)$ is well-defined for any given pair $(\mathbf{x}, y)$. We consider the two cases where the score $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)$ is non-zero and zero. In the former case, Theorem 4 shows that the sequence $\sqrt{n} \cdot [\ell(\hat{\boldsymbol{\theta}}_n; \mathbf{x}, y) - \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)]$

is asymptotically distributed according to a zero-mean normal distribution with variance shown in (17). On the other hand, when the score is zero, the rate of convergence of log-likelihood ratio is shown to be one. More specifically, Theorem 4 derives the asymptotic (non-degenerate) distribution of the sequence $n \cdot [\ell(\hat{\boldsymbol{\theta}}_n; \mathbf{x}, y) - \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)]$. Based on Proposition 26, any sequence of random variables converging in law to a non-degenerate distribution is of $O_p(1)$, that is, $n \cdot [\ell(\hat{\boldsymbol{\theta}}_n; \mathbf{x}, y) - \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)] = O_p(1)$ and therefore, $\sqrt{n} \cdot [\ell(\hat{\boldsymbol{\theta}}_n; \mathbf{x}, y) - \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)] = O_p\left(\frac{1}{\sqrt{n}}\right)$. Now, from Proposition 25, one can conclude that the sequence $\sqrt{n} \cdot [\ell(\hat{\boldsymbol{\theta}}_n; \mathbf{x}, y) - \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)]$ is of $o_p(1)$ and, by definition, converges in probability to zero. In other words, in case of zero score, the sequence $\sqrt{n} \cdot [\ell(\hat{\boldsymbol{\theta}}_n; \mathbf{x}, y) - \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)]$ converges in law to a degenerate distribution concentrated at zero and hence has zero asymptotic variance.

According to the analysis above, we can compute the expected value of the asymptotic variance by only considering regions with non-zero score. Define the region $R_0 \subseteq X \times \{1, ..., c\}$ as

$$R_0 \;:=\; \{(\mathbf{x}, y) | \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) = 0\}. \tag{22}$$

By this definition, the case of having zero score, $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) = 0$, happens with probability $P_{\boldsymbol{\theta}_0}(R_0)$, and non-zero score, $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) \neq 0$, happens with probability $1 - P_{\boldsymbol{\theta}_0}(R_0)$. Considering these two cases and the analysis made above, variance of the asymptotic distribution of $\sqrt{n} \cdot [\ell(\hat{\boldsymbol{\theta}}_n; \mathbf{x}, y) - \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)]$ can be written as

$$\begin{aligned}
&\mathrm{Var}\left(\lim_{n\to\infty} \sqrt{n} \cdot [\ell(\hat{\boldsymbol{\theta}}_n; \mathbf{x}, y) - \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)]\right) \\
&\quad = \; [1 - P_{\boldsymbol{\theta}_0}(R_0)] \cdot \mathrm{tr}\left[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) \cdot \nabla_{\boldsymbol{\theta}}^\top \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) \cdot \mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}\right] \;+\; P_{\boldsymbol{\theta}_0}(R_0) \cdot 0 \\
&\quad \leq \; \mathrm{tr}\left[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) \cdot \nabla_{\boldsymbol{\theta}}^\top \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) \cdot \mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}\right].
\end{aligned} \tag{23}$$

Taking the expectation of both sides with respect to the true joint gives the inequality (21). If the set of pairs $(\mathbf{x}, y)$ where $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) = \mathbf{0}$ form a zero measure set under $P_{\boldsymbol{\theta}_0}$, then $P_{\boldsymbol{\theta}_0}(R_0) = 0$ and we get equality in (23), hence an equality in (21). ■

Theorem 5 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.

## 5. Fisher Information Ratio in Practice

In this section, we explain how inequality (21) can be utilized in practice as an objective function for active learning. The left-hand-side is the objective that is more reasonable to minimize from classification point of view. However, its optimization is intractable and FIR-based methods approximate it by its upper-bound minimization. Querying can be done with this objective by first learning the optimal proposal distribution $q$ that minimizes the left-hand-side of inequality (21) and then drawing the queries from this optimal distribution:

$$q^* \;=\; \arg\min_q \;\; \mathrm{tr}[\mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}\, \mathbf{I}_p(\boldsymbol{\theta}_0)] \tag{24a}$$

$$X_q \;\sim\; q^*(\mathbf{x}), \tag{24b}$$

where $X_q$ is the set of queries whose samples are drawn from $q^*$. Note that in (24), due to the sampling process, $X_q$ cannot be deterministically determined even by fixing all the other parameters leading to a fixed query distribution $q^*$ (ignoring the uncertainties in the numerical optimization processes). Hence, this setting is sometimes called *probabilistic* active learning. Notice that in pool active learning, $q$ should be constrained to be a PMF over the unlabeled pool from which the queries are to be chosen. Relaxing $q$ to continuous distributions leads to *synthetic* active learning, since each time an unseen sample will be *synthesized* by sampling from $q$. We will see later that in some pool-based applications, the *objective functional* of $q$ is approximated as a *set function* of $X_q$, and therefore a combinatorial optimization is performed directly with respect to the query set.

As mentioned, $q^*$ is an upper-bound minimization of the expected asymptotic loss variance. Moreover, there are a number of unknown variables involved in FIR objective, such as $\mathbf{I}_p$ and $\boldsymbol{\theta}_0$. In practice, estimations of these unknown variables are used in the optimization process for active learning. Therefore, although the derivations in the previous section (Theorem 5) are made based on one querying of infinitely many samples, in active learning a finite-sample approximation of the cost function is used in an iterative querying process. As the number of querying iterations in active learning increases, the parameter estimates get more accurate and so does the approximate FIR objective. In the next section, we show that under certain assumptions the optimization with respect to proposal distribution in each iteration is yet another upper-bound minimization similar to (21). More specifically, Remark 6 (see Section 5.1.1) shows that although the proposal distribution is optimized separately in each iteration of an FIR-based active learning algorithm, minimizing the approximate FIR at each iteration is still an upper-bound minimization of the original cost function (i.e. left-hand-side of inequality 21).

Algorithm 0 shows steps of a general discriminative classification with active learning. We assume an initial training set $\mathcal{L}_{n_0} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_{n_0}, y_{n_0})\}$ is given based on which an initial MLE $\hat{\boldsymbol{\theta}}_{n_0}$ 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 $\mathbf{x} \in X_q$, we request its labels $y(\mathbf{x})$ from the oracle, or equivalently, sample it from the true conditional, $y(\mathbf{x}) \sim p(y|\,\mathbf{x}, \boldsymbol{\theta}_0)$. These pairs are then added into the training set to get $\mathcal{L}_{n_1}$, which in turn, is used to update the parameter estimate to $\hat{\boldsymbol{\theta}}_{n_1}$. 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 $i_{max}$. All different techniques that we discuss in the following sections, differ only in line 3 and the rest of the steps are common between them. Each active learning algorithm $\mathcal{A}$ takes the current estimate of the parameter $\hat{\boldsymbol{\theta}}_{n_{i-1}}$ possibly together with the unlabeled set of samples $X_p$, and generate a set of queries $X_q$ to be labeled for the next iteration.

In our analysis in the subsequent sections, we focus on a specific *querying iteration* indexed by $i$ (as a positive integer). For simplicity, we replace $n_{i-1}$ and $n_i$ (size of the training data set before and after iteration $i$) by $n'$ and $n$, respectively. Hence, iteration $i$ consists of using the available parameter estimate, $\hat{\boldsymbol{\theta}}_{n'}$, obtained through the current training data set $\mathcal{L}_{n'}$, to generate queries using a given querying algorithm $\mathcal{A}(\hat{\boldsymbol{\theta}}_{n'}, \mathcal{L}_{n'})$, and then update the classifier's parameter estimate accordingly to $\hat{\boldsymbol{\theta}}_n$.

In what follows, we first discuss practical issues in using FIR in query selection (Section 5.1) and then review existing algorithms based on this objective (Section 5.2).

**Algorithm 0:** Classification with Active Learning

**Inputs:** The initial training set $\mathcal{L}_{n_0}$; number of querying iterations $i_{max}$
**Outputs:** The trained classifier with MLE $\hat{\boldsymbol{\theta}}_{n_{i_{max}}}$

/* Initializations */
$\hat{\boldsymbol{\theta}}_{n_0} \leftarrow \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathcal{L}_{n_0})$
/* Starting the Iterations */
**for** $i = 1 \rightarrow i_{max}$ **do**
  /* Generating the query set by optimizing a querying objective */
  $X_q \leftarrow \mathcal{A}(\mathcal{L}_{n_{i-1}}, \hat{\boldsymbol{\theta}}_{n_{i-1}})$
  /* Request the queries' labels from the oracle */
  $y(\mathbf{x}) \sim p(y|\,\mathbf{x}, \boldsymbol{\theta}_0) \quad \forall\, \mathbf{x} \in X_q$
  /* Taking care of indexing */
  $n_i \leftarrow n_{i-1} + |X_q|$
  /* Update the training set and update MLE */
  $\mathcal{L}_{n_i} \leftarrow \mathcal{L}_{n_{i-1}} \cup \left\{ \bigcup_{\mathbf{x} \in X_q} (\mathbf{x}, y(\mathbf{x})) \right\}$
  $\hat{\boldsymbol{\theta}}_{n_i} \leftarrow \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathcal{L}_{n_i})$
**return** $\hat{\boldsymbol{\theta}}_{n_{i_{max}}}$

### 5.1 Practical Issues

The main difficulties consist of (1) having unknown variables in the objective, such as the test marginal, $p(\mathbf{x})$, and the true parameter, $\boldsymbol{\theta}_0$, and (2) lack of closed form for Fisher information matrices for most cases. In the next two sections, we review different hacks and solutions that have been proposed to resolve these issues.

#### 5.1.1 Replacing $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}}_{n'}$

Since $\boldsymbol{\theta}_0$ is not known, the simplest idea is to replace it by the current parameter estimate, that is $\hat{\boldsymbol{\theta}}_{n'}$ (Fukumizu, 2000; Settles and Craven, 2008; Hoi et al., 2006, 2009; Chaudhuri et al., 2015b). Clearly, as the algorithm keeps running the iterations ($n'$ increases), the approximate objective (which contains $\boldsymbol{\theta}_{n'}$ instead of $\boldsymbol{\theta}_0$) gets closer to the original objective. This is due to the regularity and invertibility conditions assumed for the log-likelihood function and Fisher information matrices, respectively. Moreover, Chaudhuri et al. (2015b) analyzed how this approximation effects the querying performance in finite-sample case.

Their analysis is done only for pool-based active learning, and when the test marginal $p(\mathbf{x})$ is a *uniform distribution* $U(\mathbf{x})$ over the pool $X_p$. It is also assumed that the Hessian $\frac{\partial^2 \ell(\boldsymbol{\theta};\mathbf{x},y)}{\partial\, \boldsymbol{\theta}^2}$ is independent of the class labels $y$, and therefore can be viewed as the conditional Fisher information $\mathbf{I}(\boldsymbol{\theta}, \mathbf{x})$, or equivalently $\mathbf{I}_p(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{x})}[\mathbf{I}(\boldsymbol{\theta}, \mathbf{x})]$. Furthermore, there assumed to exist four positive constants $L_1, L_2, L_3, L_4 \geq 0$ such that the following four

inequalities hold for all $\mathbf{x} \in X_p$, $y \in \{1, ..., c\}$ and $\boldsymbol{\theta} \in \Theta$:

$$\begin{aligned} \nabla \ell(\boldsymbol{\theta}_0; \mathbf{x}, y)^\top \, \mathbf{I}_p(\boldsymbol{\theta}_0)^{-1} \nabla \ell(\boldsymbol{\theta}_0; \mathbf{x}, y) &\leq L_1 \\ \left\| \mathbf{I}_p(\boldsymbol{\theta}_0)^{-1/2} \, \mathbf{I}(\boldsymbol{\theta}_0, \mathbf{x}) \, \mathbf{I}_p(\boldsymbol{\theta}_0)^{-1/2} \right\| &\leq L_2 \\ \left\| \mathbf{I}_p(\boldsymbol{\theta}_0)^{-1/2} (\mathbf{I}(\boldsymbol{\theta}', \mathbf{x}) - \mathbf{I}(\boldsymbol{\theta}'', \mathbf{x})) \, \mathbf{I}_p(\boldsymbol{\theta}_0)^{-1/2} \right\| &\leq L_3 (\boldsymbol{\theta}' - \boldsymbol{\theta}'')^\top \, \mathbf{I}_p(\boldsymbol{\theta}_0) (\boldsymbol{\theta}' - \boldsymbol{\theta}'') \\ -L_4 \| \, \boldsymbol{\theta} - \boldsymbol{\theta}_0 \, \|_2 \, \mathbf{I}(\boldsymbol{\theta}_0, \mathbf{x}) \;\preceq\; \mathbf{I}(\boldsymbol{\theta}, \mathbf{x}) - \mathbf{I}(\boldsymbol{\theta}_0, \mathbf{x}) &\preceq L_4 \| \, \boldsymbol{\theta} - \boldsymbol{\theta}_0 \, \|_2 \, \mathbf{I}(\boldsymbol{\theta}_0, \mathbf{x}), \end{aligned} \tag{25}$$

where $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}''$ are any two parameters in a fixed neighborhood of $\boldsymbol{\theta}_0$. Then, provided that $n'$ is large enough, the following remark can be shown regarding the relationship between FIRs computed at $\boldsymbol{\theta}_0$ and an estimate $\hat{\boldsymbol{\theta}}_{n'}$:

**Remark 6** *Let the assumptions (A0) to (A8) and those in (25) hold. Moreover, assume that the Hessian is independent of the class labels. If $n'$ is large enough, then the following inequality holds for any $\beta \geq 10$ with high probability:*

$$\operatorname{tr}\left[\mathbf{I}_q(\boldsymbol{\theta}_0)^{-1} \, \mathbf{I}_p(\boldsymbol{\theta}_0)\right] \;\leq\; \frac{\beta + 1}{\beta - 1} \cdot \operatorname{tr}\left[\mathbf{I}_q(\hat{\boldsymbol{\theta}}_{n'})^{-1} \, \mathbf{I}_p(\hat{\boldsymbol{\theta}}_{n'})\right]. \tag{26}$$

*The minimum value for $n'$ that is necessary for having this inequality with probability $1 - \delta$, increases quadratically with $\beta$ and reciprocally with $\delta$ (Chaudhuri et al., 2015b, Lemma 2).*

**Proof** It is shown in the proof of Lemma 2 in Chaudhuri et al. (2015a) that under assumptions mentioned in the statement, the following inequalities hold with probability $1 - \delta$:

$$\frac{\beta - 1}{\beta} \, \mathbf{I}(\mathbf{x}, \boldsymbol{\theta}_0) \;\preceq\; \mathbf{I}(\mathbf{x}, \hat{\boldsymbol{\theta}}_{n'}) \;\preceq\; \frac{\beta + 1}{\beta} \, \mathbf{I}(\mathbf{x}, \boldsymbol{\theta}_0). \tag{27}$$

Taking expectation with respect to $p(\mathbf{x})$ and $q(\mathbf{x})$ results

$$\frac{\beta - 1}{\beta} \, \mathbf{I}_p(\boldsymbol{\theta}_0) \;\preceq\; \mathbf{I}_p(\hat{\boldsymbol{\theta}}_{n'}) \;\preceq\; \frac{\beta + 1}{\beta} \, \mathbf{I}_p(\boldsymbol{\theta}_0), \tag{28a}$$

$$\frac{\beta - 1}{\beta} \, \mathbf{I}_q(\boldsymbol{\theta}_0) \;\preceq\; \mathbf{I}_q(\hat{\boldsymbol{\theta}}_{n'}) \;\preceq\; \frac{\beta + 1}{\beta} \, \mathbf{I}_q(\boldsymbol{\theta}_0). \tag{28b}$$

Since $\mathbf{I}_q(\boldsymbol{\theta}_0)$ and $\mathbf{I}_q(\hat{\boldsymbol{\theta}}_{n'})$ are assumed to be positive definite, we can write (28b) in terms of inverted matrices:[5]

$$\frac{\beta}{\beta + 1} \, \mathbf{I}_q^{-1}(\boldsymbol{\theta}_0) \;\preceq\; \mathbf{I}_q^{-1}(\hat{\boldsymbol{\theta}}_{n'}) \;\preceq\; \frac{\beta}{\beta - 1} \, \mathbf{I}_q^{-1}(\boldsymbol{\theta}_0). \tag{29}$$

Now considering the first inequalities of (28a) and (29), multiplying both sides and taking the trace result (26). ■

Inequality (26) implies that minimizing $\operatorname{tr}\left[\mathbf{I}_q(\hat{\boldsymbol{\theta}}_{n'})^{-1} \, \mathbf{I}_p(\hat{\boldsymbol{\theta}}_{n'})\right]$ (or an approximation of it) with respect to $q$ in each iteration of FIR-based querying algorithms, namely through the operation $\mathcal{A}(\mathcal{L}_{n'}, \hat{\boldsymbol{\theta}}_{n'})$ (line 3 of Algorithm 0), is equivalent to upper bound minimization of the original cost function, i.e. left-hand-side of (21).

5. For any two positive definite matrices $\mathbf{A}$ and $\mathbf{B}$, we have that $\mathbf{A} \succeq \mathbf{B} \;\Rightarrow\; \mathbf{A}^{-1} \preceq \mathbf{B}^{-1}$.

### 5.1.2 Monte-Carlo Approximation

Computation of Fisher information matrices is intractable unless when the marginal distributions are very simple or when they are restricted to be PMFs over finite number of samples. The latter is widely used in pool-based active learning, when the samples in the pool are assumed to be generated from $p(\mathbf{x})$. In such cases, one can simply utilize a Monte-Carlo approximation to compute $\mathbf{I}_p$. More specifically, denote the set of observed instances in the pool by $X_p$. Then the test Fisher information at any $\boldsymbol{\theta} \in \Omega$ can be approximated by

$$\mathbf{I}_p(\boldsymbol{\theta}) \;\approx\; \hat{\mathbf{I}}(\boldsymbol{\theta}; X_p) \;:=\; \frac{1}{|X_p|} \sum_{\mathbf{x} \in X_p} \sum_{y=1}^{c} p(y|\,\mathbf{x}, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}, y) \nabla_{\boldsymbol{\theta}}^{\top} \ell(\boldsymbol{\theta}; \mathbf{x}, y) \;+\; \delta \cdot \mathbb{I}_d, \quad (30)$$

where $\delta$ is a small positive number and the weighted identity matrix is added to ensure positive definiteness. It is important to give the practical remark that when using equation (30), we are actually using 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.

Similarly, $\mathbf{I}_q$ can be estimated based on a candidate query set $X_q$. Let $X_q$ be the set of samples drawn independently from $q(\mathbf{x})$. Then, for any $\boldsymbol{\theta} \in \Theta$, we can have the approximation $\mathbf{I}_q(\boldsymbol{\theta}) \approx \hat{\mathbf{I}}(\boldsymbol{\theta}; X_q)$. Putting everything together, the best query set $X_q \subseteq X_p$ is chosen to be the one that minimizes the approximate FIR querying objective,

$$\operatorname{tr}\left[\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}}_{n'}; X_q)^{-1} \hat{\mathbf{I}}(\hat{\boldsymbol{\theta}}_{n'}; X_p)\right]. \quad (31)$$

Note that this objective is directly written in terms of $X_q$, and therefore the queries can be deterministically determined by fixing all the rest (including the current parameter estimate $\hat{\boldsymbol{\theta}}_{n'}$) and optimizing with respect to $X_q$. Therefore, such settings are usually called *deterministic* active learning, as opposed to the probabilistic nature of (24).

### 5.1.3 Bound Optimization

There are other types of approximation methods occurring in the optimization side. These methods are able to remove part of the unknown variables by doing upper-bound minimization or lower-bound maximization. Recall that in active learning, the querying objective is to be optimized with respect to $q$ (or $X_q$ in pool-based scenario). In a very simple example, when $d = 1$, note that the $\mathbf{I}_p(\boldsymbol{\theta}_0)$ is a constant scalar in (24a) and hence can be ignored. 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 $\mathbf{I}_q(\hat{\boldsymbol{\theta}}_n)$ to optimize, one may choose the objective to be $|\,\mathbf{I}_q(\hat{\boldsymbol{\theta}}_n)|$ (where $|\cdot|$ is the determinant function),[6] or $\operatorname{tr}[\mathbf{I}_q(\hat{\boldsymbol{\theta}}_n)]$.[7] The latter is worth paying more attention due to the following inequality (Yang, 2000):

$$\operatorname{tr}[\mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}\, \mathbf{I}_p(\boldsymbol{\theta}_0)] \;\leq\; \operatorname{tr}[\mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}] \cdot \operatorname{tr}[\mathbf{I}_p(\boldsymbol{\theta}_0)]. \quad (32)$$

Since $\operatorname{tr}[\mathbf{I}_p(\boldsymbol{\theta}_0)]$ is a constant with respect to $q$, minimizing the right-hand-side of inequality (21) can itself be approximated by another upper-bound minimization:

$$\arg\min_{q} \;\; \operatorname{tr}[\mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}]. \quad (33)$$

6. Similar to *D-optimality* in Optimal Experiment Design (Fedorov, 1972).
7. Similar to *A-optimality* in Optimal Experiment Design (Fedorov, 1972).

| | Algorithm | Obj. | Prob. | Det. | Pool | Syn. | Seq. | Batch |
|---|---|---|---|---|---|---|---|---|
| 1 | Fukumizu (2000) | (34) | ✓ | | | ✓ | ✓ | ✓ |
| 2 | Zhang and Oles (2000) | (34) | | ✓ | ✓ | ✓ | ✓ | |
| 3 | Settles and Craven (2008) | (31) | | ✓ | ✓ | | ✓ | ✓ |
| 4 | Hoi et al. (2006, 2009) | (31) | | ✓ | ✓ | | ✓ | ✓ |
| 5 | Chaudhuri et al. (2015b) | (26) | ✓ | | ✓ | | ✓ | ✓ |

Table 1: Reviewed FIR-based active learning algorithms for discriminative classifiers

This helps removing the dependence of the objective to the test distribution $p$. A lower bound can also be established for the FIR. Using the inequality between arithmetic and geometric means of the eigenvalues of $\mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}\,\mathbf{I}_p(\boldsymbol{\theta}_0)$, one can see that $d \cdot |\,\mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}| \cdot |\,\mathbf{I}_p(\boldsymbol{\theta}_0)| \leq \mathrm{tr}[\mathbf{I}_q(\boldsymbol{\theta}_0)^{-1}\,\mathbf{I}_p(\boldsymbol{\theta}_0)]$. Hence, when minimizing the upper-bound in (33), 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

$$\arg\max_{q} \;\; \mathrm{tr}[\mathbf{I}_q(\boldsymbol{\theta}_0)]. \tag{34}$$

Hence, algorithms that aim to maximize $\mathrm{tr}[\mathbf{I}_q(\boldsymbol{\theta}_0)]$, indeed introduce two layers of objective approximations through equations (32) to (34). As discussed before, the dependence of the objectives in all the layers (in 33 or 34) on $\boldsymbol{\theta}_0$ can be removed by replacing it with the current estimate $\boldsymbol{\theta}_{n'}$.

### 5.2 Some Existing Algorithms

In this section, we discuss several existing algorithms for implementing the query selection task based on minimization of FIR. We will analyze these algorithms, sorted according to date of their publication, in the context of our unifying framework.

Besides the categorizations that have already been described in previous sections, 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.

Table 1 lists the algorithms that we reviewed in the following sections together with a summary of their properties and the approximate objective that they optimize for querying. Note that among these algorithms, the one by Chaudhuri et al. (2015b) makes extra assumptions as is described in Section 5.1.1.

*Algorithm 1 (Fukumizu, 2000)*

This algorithm is the classification version of the *probabilistic active learning* proposed by Fukumizu (2000) for regression problem. The assumption is that the proposal belongs to a parametric family and is of the form $q(\mathbf{x};\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is the parameter vector of the

**Algorithm 1:** Fukumizu (2000)

**Inputs:** Current estimation of the parameter $\hat{\boldsymbol{\theta}}_{n'}$, size of the query set $|X_q|$
**Outputs:** The query set $X_q$

/* Parameter Optimization */
$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha})} \left[ \sum_{y=1}^{c} p(y|\,\mathbf{x}, \hat{\boldsymbol{\theta}}_{n'}) \nabla_{\boldsymbol{\theta}}^{\top} \ell(\hat{\boldsymbol{\theta}}_{n'}; \mathbf{x}, y) \nabla_{\boldsymbol{\theta}} \ell(\hat{\boldsymbol{\theta}}_{n'}; \mathbf{x}, y) \right]$
/* Sampling from the parametric proposal */
$\mathbf{x}_i \sim q(\mathbf{x}; \hat{\boldsymbol{\alpha}}) \quad , i = 1, ..., |X_q|$
**return** $X_q = \left\{ \mathbf{x}_1, ..., \mathbf{x}_{|X_q|} \right\}$

family. In this *parametric* active learning, the best set of parameters $\hat{\boldsymbol{\alpha}}$ is selected using the current parameter estimate and the query set is sampled from the resulting proposal distribution $X_q \sim q(\mathbf{x}; \hat{\boldsymbol{\alpha}})$.

This algorithm makes no use of the test samples and optimizes the simplified objective in (34) to obtain the query distribution $q(\mathbf{x})$. Denote the covariates of the current training data set $\mathcal{L}_{n'}$ by $X_{\mathcal{L}}$. As is described in Section (5.1), the new trace objective can be approximated by Monte-Carlo formulation using the old queried samples $X_{\mathcal{L}}$ as well as the candidate queries $X_q$ to be selected in this iteration:

$$\operatorname{tr}\left[ \hat{\mathbf{I}}(\hat{\boldsymbol{\theta}}_{n'}; X_{\mathcal{L}} \cup X_q) \right]. \tag{35}$$

More specifically, the new parameter vector is obtained by maximizing the expected contribution of the queries $X_q$ generated from $q(\mathbf{x}; \boldsymbol{\alpha})$ to this objective. Taking expectation of (35) with respect to $q(\mathbf{x}; \boldsymbol{\alpha})$ yields

$$\mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha})} \left[ \operatorname{tr}\left[ \hat{\mathbf{I}}(\hat{\boldsymbol{\theta}}_{n'}; X_{\mathcal{L}} \cup X_q) \right] \right] = \operatorname{tr}\left[ \frac{n'}{n' + |X_q|} \hat{\mathbf{I}}(\hat{\boldsymbol{\theta}}_{n'}; X_{\mathcal{L}}) + \frac{1}{n' + |X_q|} \mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha})} \left[ \hat{\mathbf{I}}(\hat{\boldsymbol{\theta}}_{n'}; X_q) \right] \right]. \tag{36}$$

Recall that $n'$ is the size of the current training data set $\mathcal{L}_{n'}$. The first term in (36) is independent of the query set $X_q$ (assuming that the size $|X_q|$ is fixed to a constant), hence we focus only on the second term in our optimization. Noting that the queries are generated independently, we can rewrite this term (excluding the constant coefficient) as

$$\mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha})} \left[ \operatorname{tr}\left[ \hat{\mathbf{I}}(\hat{\boldsymbol{\theta}}_{n'}; X_q) \right] \right] = \frac{1}{|X_q|} \mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha})} \left[ \sum_{\mathbf{x} \in X_q} \operatorname{tr}\left[ \hat{\mathbf{I}}(\hat{\boldsymbol{\theta}}_{n'}; \mathbf{x}) \right] \right] \tag{37}$$

From equation (37), the parameter vector $\boldsymbol{\alpha}$ can be obtained by maximizing the expected contribution of that single query to the trace objective. That is, having fixed $|X_q|$ to a constant, the optimization for determining the parameter vector would be

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= \arg\max_{\boldsymbol{\alpha}} \; \mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha})} \left[ \operatorname{tr}\left[ \hat{\mathbf{I}}(\hat{\boldsymbol{\theta}}_{n'}; \mathbf{x}) \right] \right] \\ &= \arg\max_{\boldsymbol{\alpha}} \; \mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha})} \left[ \sum_{y=1}^{c} p(y|\,\mathbf{x}, \hat{\boldsymbol{\theta}}_{n'}) \nabla_{\boldsymbol{\theta}}^{\top} \ell(\hat{\boldsymbol{\theta}}_{n'}; \mathbf{x}, y) \nabla_{\boldsymbol{\theta}} \ell(\hat{\boldsymbol{\theta}}_{n'}; \mathbf{x}, y) \right]. \end{aligned} \tag{38}$$

**Algorithm 2:** Zhang and Oles (2000)

**Inputs:** Current estimation of the parameter $\hat{\boldsymbol{\theta}}_{n'}$
**Outputs:** The query singleton set $X_q = \{\mathbf{x}_q\}$

$\mathbf{x}_q \leftarrow \arg\max_{\mathbf{x}} \sum_{y=1}^{c} p(y|\,\mathbf{x}, \hat{\boldsymbol{\theta}}_{n'}) \nabla_{\boldsymbol{\theta}}^{\top} \ell(\hat{\boldsymbol{\theta}}_{n'}; \mathbf{x}, y) \nabla_{\boldsymbol{\theta}} \ell(\hat{\boldsymbol{\theta}}_{n'}; \mathbf{x}, y)$
**return** $X_q = \{\mathbf{x}_q\}$

The optimization (38) does not depend on $X_{\mathcal{L}}$, and therefore we do not need to explicitly feed this algorithm with $\mathcal{L}$. All it needs is an estimation of the parameter $\hat{\boldsymbol{\theta}}_{n'}$. The two-step procedure of generating queries from parametric query distribution is shown in Algorithm 1. This algorithm can be used in both sequential and batch modes by changing the number of samples drawn from $q(\mathbf{x}; \alpha)$.

We emphasize that Algorithm 1 is probabilistic, meaning that with any fixed parameter estimate $\hat{\boldsymbol{\theta}}_{n'}$, 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 (2000) 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.

*Algorithm 2 (Zhang and Oles, 2000)*

Zhang and Oles (2000) started from optimization problem (34), and introduced even additional simplifications to it. They specifically considered using 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 $X_q = \{\mathbf{x}_q\}$. Similar to the previous section, the Fisher information matrix $\mathbf{I}_q$ can be approximated by Monte-Carlo approximation. Zhang and Oles (2000) discarded the expectation with respect to the proposal distribution in (38) 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 $\mathbf{x}_q$:

$$\mathbf{x}_q \;=\; \arg\max_{\mathbf{x} \in X} \; \sum_{y=1}^{c} p(y|\,\mathbf{x}, \hat{\boldsymbol{\theta}}_{n'}) \nabla_{\boldsymbol{\theta}}^{\top} \ell(\hat{\boldsymbol{\theta}}_{n'}; \mathbf{x}, y) \nabla_{\boldsymbol{\theta}} \ell(\hat{\boldsymbol{\theta}}_{n'}; \mathbf{x}, y). \tag{39}$$

This single-step 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 $\mathbf{x}_q$ to be a member of a pool of samples, in which case it can even be extended to batch querying by sorting the unlabeled samples based on their objective values and taking the highest ones. 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 (2008)

**Inputs:** Current estimation of the parameter $\hat{\boldsymbol{\theta}}_{n'}$, the set of unlabeled samples $X_p$, , size of the query set $|X_q|$

**Outputs:** The query set $X_q$

```
/* Initializing the query set for this iteration                          */
```
$X_q \leftarrow \varnothing$

```
/* The loop for greedy batch querying                                     */
```
**for** $j = 1 \rightarrow |X_q|$ **do**

```
    /* Query optimization and adding the result into the query set        */
```
    $X_q \leftarrow X_q \cup \arg\min_{\mathbf{x} \in X_p} \operatorname{tr}\left[\hat{\mathbf{I}}\left(\hat{\boldsymbol{\theta}}_{n'}; \mathbf{x}\right)^{-1} \hat{\mathbf{I}}\left(\hat{\boldsymbol{\theta}}_{n'}; X_p\right)\right]$

```
    /* Removing the selected queries from the pool                        */
```
    $X_p \leftarrow X_p - X_q$

**return** $X_q$

*Algorithm 3 (Settles and Craven, 2008)*

Inspired by Zhang and Oles (2000), Settles and Craven (2008) employed Fisher information ratio to develop a pool-based active learning, which can be used in either sequential or batch querying. The pool that is used here is the set of unlabeled samples, $X_p$, which are assumed to be drawn from the test marginal $p(\mathbf{x})$. Queries are chosen from $X_p$, 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 $\hat{\mathbf{I}}\left(\hat{\boldsymbol{\theta}}_{n'}; X_p\right)$. Similar to Algorithm 1, the updated training Fisher information matrix after querying a set $X_q$ can be approximated by $\hat{\mathbf{I}}\left(\hat{\boldsymbol{\theta}}_{n'}; X_{\mathcal{L}} \cup X_q\right)$. 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 (2000) algorithm, the proposal distribution $q$ is ignored in the objective (or equivalently considered as being uniform). An additional assumption Settles and Craven (2008) made to simplify the optimization task is

$$\underset{X_q \subset X_p}{\arg\min}\ \operatorname{tr}\left[\hat{\mathbf{I}}\left(\hat{\boldsymbol{\theta}}_{n'}; X_{\mathcal{L}} \cup X_q\right)^{-1} \hat{\mathbf{I}}\left(\hat{\boldsymbol{\theta}}_{n'}; X_p\right)\right] \approx \underset{X_q \subset X_p}{\arg\min}\ \operatorname{tr}\left[\hat{\mathbf{I}}\left(\hat{\boldsymbol{\theta}}_{n'}; X_q\right)^{-1} \hat{\mathbf{I}}\left(\hat{\boldsymbol{\theta}}_{n'}; X_p\right)\right]. \tag{40}$$

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 (2008) used a greedy approach to do this optimization (the inner loop).

*Algorithm 4 (Hoi et al., 2006, 2009)*

The algorithms proposed by Hoi et al. (2006) and Hoi et al. (2009) are very similar to the one developed by Settles and Craven (2008), described above, except that they use a more sophisticated optimization method. Their method shown in Algorithm 4, is different

**Algorithm 4:** Hoi et al. (2006, 2009)

**Inputs:** Current estimation of the parameter $\hat{\boldsymbol{\theta}}_{n'}$, the set of unlabeled samples $X_p$, size of the query set $|X_q|$

**Outputs:** The query set $X_q$

/* Initializing the query set */

1 $X_q \leftarrow \varnothing$

/* The loop for greedy batch querying */

2 **for** $j = 1 \rightarrow |X_q|$ **do**

/* Query optimization */

3 $\tilde{\mathbf{x}} = \arg\min_{\mathbf{x} \in X_p} \operatorname{tr}\left[\hat{\mathbf{I}}\left(\hat{\boldsymbol{\theta}}_{n'}; X_q \cup \{\mathbf{x}\}\right)^{-1} \hat{\mathbf{I}}\left(\hat{\boldsymbol{\theta}}_{n'}; X_p\right)\right]$

/* Add the selected query into the query set */

4 $X_q \leftarrow X_q \cup \{\tilde{\mathbf{x}}\}$

/* Remove the selected instance from the pool */

5 $X_p \leftarrow X_p - \{\tilde{\mathbf{x}}\}$

6 **return** $X_q$

from Algorithm 3 mainly in the way that it greedily chooses the query at each inner loop iteration of the algorithm. While Algorithm 3 exclusively considers the contribution of each $\mathbf{x} \in X_q$, ignoring the samples selected in the previous iterations and considering only $\hat{\mathbf{I}}\left(\hat{\boldsymbol{\theta}}_{n'}; \mathbf{x}\right)$ in each iteration (line 3 of Algorithm 3), Algorithm 4 takes into account all the queries chosen so far and instead considers $\hat{\mathbf{I}}\left(\hat{\boldsymbol{\theta}}_{n'}; X_q \cup \{\mathbf{x}\}\right)$ in each querying optimization (line 3 in Algorithm 4).

Hoi et al. (2006) and Hoi et al. (2009) showed that when using *binary logistic regression* classifier, their optimization (40) can be done by maximizing a *submodular* set function with respect to the query set $X_q$. This allowed them to use the well-known iterative algorithm proposed by Nemhauser et al. (1978), 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. (2006) that is obtained in case of using logistic regression classifier. As a consequence, FIR can be efficiently optimized with guaranteed tight bounds (Nemhauser et al., 1978; Nemhauser and Wolsey, 1978). First, we show in the following lemma that (40) is approximately equivalent to maximizing a simplified set function, for any unlabeled sample pool $X_p$:

**Lemma 7** *Let $X_p, X_q \subseteq X$ be two non-empty and finite sets of samples randomly generated from $p(\mathbf{x})$ and its resample distribution $q(\mathbf{x})$, respectively, such that $X_q \subset X_p$, 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 $g_{\boldsymbol{\theta}} : X \times \{1, ..., c\} \times X \rightarrow \mathbb{R}^+$, d-dimensional non-zero vector $\mathbf{v}_{\boldsymbol{\theta}}$ depending on $\mathbf{x}$ and $y$, and*

*for all* $\boldsymbol{\theta} \in \Omega$ *:*

$$(i) \quad \underset{X_q \subset X_p}{\arg\min} \ \operatorname{tr}\left[\hat{\mathbf{I}}(\boldsymbol{\theta}; X_q)^{-1} \hat{\mathbf{I}}(\boldsymbol{\theta}; X_p)\right], \tag{41a}$$

$$(ii) \quad \underset{X_q \subset X_p}{\arg\max} \sum_{\mathbf{x} \in X_p - X_q} \sum_{y=1}^{c} \frac{-1}{\delta \cdot \| \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}, y)\|^{-2} + \sum_{\mathbf{x}' \in X_q} g_{\boldsymbol{\theta}}(\mathbf{x}, y, \mathbf{x}')}. \tag{41b}$$

*The approximation is more accurate for smaller* $\delta$ *and well-conditioned Monte-Carlo approximation of proposal Fisher information matrix.*

The proof can be found in Appendix C. Note that Lemma 7, 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 8** *Suppose* $f_{\boldsymbol{\theta}} : 2^{X_p} \to \mathbb{R}$ *is defined as*

$$f_{\boldsymbol{\theta}}(X_q) \ = \ \sum_{\mathbf{x} \in X_p - X_q} \sum_{y=1}^{c} \frac{-1}{\delta \cdot \| \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}, y)\|^{-2} + \sum_{\mathbf{x}' \in X_q} g_{\boldsymbol{\theta}}(\mathbf{x}, y, \mathbf{x}')}, \quad \forall X_q \subseteq X_p, \tag{42}$$

*with* $\mathbf{v}_{\boldsymbol{\theta}}$ *a* $d$*-dimensional vector depending on* $\mathbf{x}$ *and* $y$*, and* $g_{\boldsymbol{\theta}}$ *defined in (91). Then* $f_{\boldsymbol{\theta}}$ *is a submodular and monotone (non-decreasing) set function for all* $\boldsymbol{\theta} \in \Omega$*.*

The proof is in Appendix D. The result above, together with Lemma 7, imply that the objective of (41b) 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. (1978) and Nemhauser and Wolsey (1978).

**Theorem 9 (Nemhauser et al. (1978))** *Let* $f_{\boldsymbol{\theta}} : 2^{X_p} \to \mathbb{R}$ *be any submodular and non-decreasing set function with* $f(\varnothing) = 0$ *satisfied.*[8] *If* $X_q$ *is the output of a greedy maximization algorithm, and* $X_q^*$ *is the optimal maximizer of* $f_{\boldsymbol{\theta}}$ *with a cardinality constraint (fixed* $|X_q|$*), then we have*

$$f_{\boldsymbol{\theta}}(X_q) \ \geq \ \left[1 - \left(\frac{|X_q| - 1}{|X_q|}\right)^{|X_q|}\right] f_{\boldsymbol{\theta}}(X_q^*) \ \geq \ \left(1 - \frac{1}{e}\right) f_{\boldsymbol{\theta}}(X_q^*). \tag{43}$$

In Algorithm 4, the inner loop (lines 2 to 5) implements the minimization in (40) greedily. We have seen above that this set minimization is approximately equivalent to maximizing a submodular and monotone set maximization, which, in turn, is shown to be efficient.

8. This can always be assumed since maximizing a general set function $f(X_q)$ is equivalent to maximizing its adjusted version $g(X_q) := f(X_q) - f(\varnothing)$, which satisfies $g(\varnothing) = 0$.

*Algorithm 5 (Chaudhuri et al., 2015b)*

This algorithm uses FIR for doing a probabilistic pool-based active learning. It has extra assumptions in comparison to our general framework, which are briefly explained in Section 5.1.1. Note that these assumptions are to be made as well as those listed in Section 2. In such settings, Chaudhuri et al. (2015b) gave a finite-sample theoretical analysis for FIR when applied to pool-based active learning.

More specifically, suppose $p(\mathbf{x})$ is a uniform PMF and $q(\mathbf{x})$ is a general PMF, both defined over the pool $X_p$. Using the notations in (25), the training Fisher information can be written as $\mathbf{I}_q(\hat{\boldsymbol{\theta}}_{n'}) = \sum_{\mathbf{x}\in X_p} q(\mathbf{x})\,\mathbf{I}(\hat{\boldsymbol{\theta}}_{n'}, \mathbf{x})$. Now, assuming that $\mathbf{I}_p(\hat{\boldsymbol{\theta}}_{n'})$ has a singular decomposition of the form $\sum_{j=1}^{d} \sigma_j \mathbf{u}_j \mathbf{u}_j^\top$, FIR can be written as

$$\begin{aligned} \mathrm{tr}\Big[\,\mathbf{I}_q(\hat{\boldsymbol{\theta}}_{n'})^{-1}\,\mathbf{I}_p(\hat{\boldsymbol{\theta}}_{n'})\Big] &= \sum_{j=1}^{d} \sigma_j \mathrm{tr}\Big[\,\mathbf{I}_q(\hat{\boldsymbol{\theta}}_{n'})^{-1}\mathbf{u}_j\mathbf{u}_j^\top\Big] \\ &= \sum_{j=1}^{d} \sigma_j \mathbf{u}_j^\top\,\mathbf{I}_q(\hat{\boldsymbol{\theta}}_{n'})^{-1}\mathbf{u}_j. \end{aligned} \tag{44}$$

Minimizing the last term in (44) with respect to PMF $\{q(\mathbf{x})|\,\mathbf{x}\in X_p\}$ is equivalent to a semidefinite programming after introducing a set of auxiliary variables $t_j, j =, 1..., d$ and applying Schur complements (Vandenberghe and Boyd, 1996):

$$\begin{aligned} \underset{q(\mathbf{x}),\mathbf{x}\in X_p}{\arg\min} \quad & \sum_{j=1}^{d} \sigma_j t_j \\ \text{such that} \quad & \begin{bmatrix} t_j & \mathbf{u}_j^\top \\ \mathbf{u}_j & \sum_{\mathbf{x}\in X_p} q(\mathbf{x})\,\mathbf{I}(\hat{\boldsymbol{\theta}}_{n'}, \mathbf{x}) \end{bmatrix} \succeq 0, \\ & \sum_{\mathbf{x}\in X_p} q(\mathbf{x}) = 1. \end{aligned} \tag{45}$$

The steps for this querying method is shown in Algorithm 5. Note that the solution to (45) is slightly modified by mixing it with the uniform distribution over the pool. Such modification is mainly to establish their theoretical derivations. The mixing coefficient, $0 \le \lambda \le 1$, reciprocally depends on the number of queries. More specifically, Chaudhuri et al. (2015b) made it equal to $1 - \frac{1}{|X_q|^{1/6}}$. That is, as the number of queries increases, $\lambda$ shrinks and so does the modification. Furthermore, in their analysis, they assumed that sampling from $\tilde{q}(\mathbf{x})$ (line 3 of Algorithm 5) is done *with replacement*. That is, label of a given sample might be queried multiple times.

### 5.2.1 Comparison with Other Information-theoretic Objectives

In the last part of this section, we compare FIR and two other common querying objectives from the field of information theory. Entropy of class labels and mutual information between labeled and unlabeled samples are two other common active learning objectives. Their goal is mainly to get the largest possible amount of information about *class labels of unlabeled samples* from each querying iteration, hence naturally pool-based.

**Algorithm 5:** Chaudhuri et al. (2015b)

**Inputs:** Current estimation of the parameter $\hat{\boldsymbol{\theta}}_{n'}$, the set of unlabeled samples $X_p$, size of the query set $|X_q|$

**Outputs:** The query set $X_q$

/* Solving the semidefinite programming */

$q(\mathbf{x}) \leftarrow$ solution to (45)

/* Modification of the solution */

$\tilde{q}(\mathbf{x}) \leftarrow \lambda q(\mathbf{x}) + (1-\lambda)U(\mathbf{x})$

/* Sampling with replacement from the modified proposal */

$\mathbf{x}_i \sim \tilde{q}(\mathbf{x}) \quad , i = 1, ..., |X_q|$

**return** $X_q = \{\mathbf{x}_1, ... \mathbf{x}_{|X_q|}\}$

| Algorithm | Complexity |
|---|---|
| Entropy | $O(\|X_p\|cd)$ |
| Mutual Information | $O\left(\|X_p\| \cdot \|X_q\| \cdot c^{\|X_q\|+1} d\right)$ |
| Zhang and Oles (2000) | $O(\|X_p\|cd)$ |
| Settles and Craven (2008) | $O\left(\|X_q\| \cdot \|X_p\| \cdot (cd + d^3)\right)$ |
| Hoi et al. (2006, 2009) | $O\left(\|X_q\| \cdot \|X_p\| \cdot (cd + cd\|X_q\| + d^3)\right)$ |
| Chaudhuri et al. (2015b) | $O\left(d^3\|X_p\|^2 + d^4\|X_p\| + d^5\right)$ |

Table 2: Computational complexity of different querying algorithms

Entropy-based querying, also known as uncertainty sampling, directly measures the uncertainty with respect to class label of each unlabeled sample and query those with highest uncertainty. It has been widely popular due to its simplicity and effectiveness especially in sequential active learning. However, it does not consider interaction between samples when selecting multiple queries, which can cause querying very similar samples (redundancy). Therefore, uncertainty sampling shows relatively poor performance in batch active learning. Mutual information, on the other hand, does not suffer from redundancy, however, it requires a much higher computational complexity.

These two objectives directly measure the amount of information each batch can have with respect to the class labels (hence prediction-based), as opposed to Fisher information as a measure of information regarding the distribution parameters (hence inference-based). However, there is no guarantee that by minimizing uncertainty of the class labels (or equivalently, choosing queries with highest amount of information about class labels), the prediction accuracy also increases. Whereas, as we showed earlier, FIR upper-bounds the expected asymptotic variance of a parameter inference loss function. From this point of view, FIR has a closer relationship with the performance of a classifier.

Table 2 shows computational complexity of the querying objectives. The algorithm by Fukumizu (2000) is excluded from this table since it cannot be used in pool-based sampling. Also note that the complexity reported for mutual information is for the case when it is optimized greedily. Nevertheless, it still contains an exponential term in its complexity. Entropy-based and Zhang and Oles (2000) have the lowest complexity, but in the expense of introducing redundancy into the batch of queries. Algorithms by Settles and Craven (2008), Hoi et al. (2006, 2009) and Chaudhuri et al. (2015b) become very expensive when $d$ is large, whereas mutual information can easily get intractable for selecting batches of higher size (large $|X_q|$). Observe that algorithm by Hoi et al. (2006, 2009) is more expensive than Settles and Craven (2008). Recall that despite similarities in appearance, the former guarantees tight bound for its greedy optimization, whereas the latter does not.

The complexity for the algorithm by Chaudhuri et al. (2015b) is computed assuming that a barrier method (following path) is used as its numerical optimization (Boyd and Vandenberghe, 2004). From Table 2, this algorithm is the only one whose complexity increases quadratically with size of the pool $|X_p|$, and therefore can get significantly slow for huge pools. Furthermore, it does not depend on $|X_q|$ since the optimization in (45) as its main source of computations, only depends on $|X_p|$ and $d$. Furthermore, for this algorithm, computing $\mathbf{I}(\hat{\boldsymbol{\theta}}_{n'}, \mathbf{x})$ is assumed to cost $O(1)$ for each $\mathbf{x} \in X_p$ as it is taken to be independent of $y$.

## 6. 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 expectation of such asymptotic variance. We also showed that replacing the true parameter by the current estimate does not remove such upper-boundedness provided that the current parameter estimate has been obtained using sufficient number of samples. Moreover, 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. Additionally, 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 existing 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.

Finally, we remark that the log-likelihood ratio that is used here as the loss function is an inference-based performance metric. It naturally shows up based on the set of assumptions that are usually being made in FIR-based querying frameworks. The final goal of a classifier in machine learning is to predict labels of test samples as accurately as possible and therefore, arguably, prediction-based metrics such as 0/1 loss function better evaluate the performance of a classifier. While analyzing such metrics was out of scope of this paper,

analysis and development of querying algorithms using prediction-based metrics is definitely an exciting future research direction.

## Acknowledgments

This work is primarily supported by NSF IIS-1118061. In addition, Todd Leen was supported under NSF 1258633, 1355603, and 1454956; Deniz Erdogmus was supported by NSF IIS-1149570, CNS-1136027 and SMA-0835976; and Jennifery Dy was also supported by NSF IIS 0915910 and NIH R01HL089856.

## Appendix A. Statistical Background

Asymptotic analysis plays an important role in statistics. It considers 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 (1998).

### A.1 Convergence of Sequence of Random Variables

Throughout this section, $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_n, ...\}$, denoted simply by $\{\boldsymbol{\theta}_n\}$, is a sequence of multivariate random variables lying in $\Omega \subseteq \mathbb{R}^d$. Also suppose that $\boldsymbol{\theta}_0$ is a constant vector and $\tilde{\boldsymbol{\theta}}$ is another random variable in the same space $\Omega$.

**Definition 10** *We say that the sequence* $\{\boldsymbol{\theta}_n\}$ converges in probability *to* $\boldsymbol{\theta}_0$ *and write* $\boldsymbol{\theta}_n \xrightarrow{P} \boldsymbol{\theta}_0$, *iff for every* $\varepsilon > 0$ *we have*

$$P(|\theta_{ni} - \theta_{0i}| > \varepsilon) \;\; \rightarrow \;\; 0, \quad \textit{for all } i = 1, ..., d, \tag{46}$$

*where* $\theta_{ni}$ *is the i'th component of* $\boldsymbol{\theta}_n$.

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

**Proposition 11 (Brockwell and Davis 1991, Proposition 6.1.4)** *If* $\boldsymbol{\theta}_n \xrightarrow{P} \boldsymbol{\theta}_0$ *and* $g : \Omega \rightarrow \mathbb{R}$ *is a continuous function, then* $g(\boldsymbol{\theta}_n) \xrightarrow{P} g(\boldsymbol{\theta}_0)$.

**Definition 12** *We say that a sequence* $\{\boldsymbol{\theta}_n\}$ converges in law (in distribution) *to the random variable* $\tilde{\boldsymbol{\theta}}$ *and write* $\boldsymbol{\theta}_n \xrightarrow{L} \tilde{\boldsymbol{\theta}}$, *iff the sequence of their joint CDFs,* $F_n$, *point-wise converges to the joint CDF of* $\tilde{\boldsymbol{\theta}}$:

$$F_n(\mathbf{a}) = P(\theta_{n1} \leq a_1, ..., \theta_{nd} \leq a_d) \;\; \rightarrow \;\; F(\mathbf{a}) = P(\tilde{\theta}_1 \leq a_1, ..., \tilde{\theta}_d \leq a_d) \quad \forall \mathbf{a} \in \mathcal{C}_F \subseteq \mathbb{R}^d, \tag{47}$$

*where* $\mathcal{C}_F$ *is the set of continuity points of the CDF* $F$.

Equation (47) means that for large values of $n$, the distribution of $\boldsymbol{\theta}_n$ can be well approximated by the distribution of $\tilde{\boldsymbol{\theta}}$. Note that throughout this paper, for simplicity, we say that a random sequence $\{\boldsymbol{\theta}_n\}$ converges to a distribution with density function $p(\boldsymbol{\theta})$, or write $\boldsymbol{\theta}_n \xrightarrow{L} p(\boldsymbol{\theta})$, instead of the full statement that $\{\boldsymbol{\theta}_n\}$ converges in law to a random variable with that distribution.

Note that $\boldsymbol{\theta}_n \xrightarrow{P} \boldsymbol{\theta}_0$ suggests that $\boldsymbol{\theta}_n - \boldsymbol{\theta}_0 \xrightarrow{L} \delta(\boldsymbol{\theta})$ where $\delta$ is the Kronecker delta function, which can be viewed as the density function of a degenerate distribution at $\boldsymbol{\theta} = \mathbf{0}$. This, however, does not give any information about the speed with which $\boldsymbol{\theta}_n$ converges to $\boldsymbol{\theta}_0$. In order to take the rate into account, we consider the convergent distribution of the sequence $\{a_n \cdot (\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)\}$, where $a_n$ is any sequence of positive integers and $a_n \to \infty$(as $n \to \infty$). In practice $a_n$ is usually considered to have the form $n^r$ with $r > 0$.

**Definition 13** *Assume $\boldsymbol{\theta}_n \xrightarrow{P} \boldsymbol{\theta}_0$. We say that the sequence $\{\boldsymbol{\theta}_n\}$ converges to $\boldsymbol{\theta}_0$ with* rate of convergence *$r > 0$, iff $n^r(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$ converges in law to a random variable with non-degenerate distribution. Furthermore, the non-degenerate distribution is called the* asymptotic distribution *of $\boldsymbol{\theta}_n$.*

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

**Theorem 14 (Law of Large Numbers, Brockwell and Davis 1991)** *Let $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n$ be a set of independent and identically distributed (i.i.d) samples. If $\mathbb{E}[\boldsymbol{\theta}_i] = \boldsymbol{\mu}$, then*

$$\bar{\boldsymbol{\theta}}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\theta}_i \;\xrightarrow{P}\; \boldsymbol{\mu}\,. \tag{48}$$

**Theorem 15 (Central Limit Theorem, Lehmann and Casella 1998)** *Let $\boldsymbol{\theta}_1$,..., $\boldsymbol{\theta}_n$ be a set of i.i.d samples with mean $\mathbb{E}[\boldsymbol{\theta}_i] = \boldsymbol{\mu}$ and covariance $\mathrm{Cov}[\boldsymbol{\theta}_i] = \boldsymbol{\Sigma}$ (with a symmetric and positive semi-definite matrix $\boldsymbol{\Sigma}$), then the sequence of sample averages $\left\{\bar{\boldsymbol{\theta}}_n\right\}$ with $\bar{\boldsymbol{\theta}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\theta}_i$ converges to the true mean with convergence rate $1/2$. Moreover, its asymptotic distribution is a zero-mean Gaussian distribution with covariance matrix $\boldsymbol{\Sigma}$, that is,*

$$\sqrt{n} \cdot (\bar{\boldsymbol{\theta}}_n - \boldsymbol{\mu}) \;\xrightarrow{L}\; \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{49}$$

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

**Theorem 16 (Multivariate Delta Method, first order, Lehmann and Casella 1998)** *Let $\{\boldsymbol{\theta}_n\}$ be a sequence of random variables such that it converges to $\boldsymbol{\theta}_0$ with rate of convergence $1/2$ and a normal asymptotic distribution, that is, $\sqrt{n} \cdot (\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. If $g : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable mapping and $\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_0) \neq \mathbf{0}$, then*

$$\sqrt{n} \cdot \Big[ g(\boldsymbol{\theta}_n) - g(\boldsymbol{\theta}_0) \Big] \;\xrightarrow{L}\; \mathcal{N}\left(0, \nabla_{\boldsymbol{\theta}}^{\top} g(\boldsymbol{\theta}_0)\, \boldsymbol{\Sigma}\, \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_0)\right). \tag{50}$$

**Theorem 17 (Multivariate Delta Method, second order)** *Let $\{\boldsymbol{\theta}_n\}$ be a sequence of random variables such that it converges to $\boldsymbol{\theta}_0$ with rate of convergence $1/2$ and a normal*

*asymptotic distribution, that is,* $\sqrt{n}\cdot(\boldsymbol{\theta}_n-\boldsymbol{\theta}_0)\xrightarrow{L}\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$*. If* $g:\mathbb{R}^d\to\mathbb{R}$ *is a continuously differentiable mapping where* $\nabla_{\boldsymbol{\theta}}g(\boldsymbol{\theta}_0)=\mathbf{0}$ *and* $\nabla^2_{\boldsymbol{\theta}}g(\boldsymbol{\theta}_0)$ *is non-singular in a neighborhood of* $\boldsymbol{\theta}_0$*, then the sequence* $\{g(\boldsymbol{\theta}_n)-g(\boldsymbol{\theta}_0)\}$ *converges in law to a mixture of random variables with degree-one Chi-square distributions, and the rate of convergence is one. More specifically,*

$$n\cdot\Big[g(\boldsymbol{\theta}_n)-g(\boldsymbol{\theta}_0)\Big]\ \xrightarrow{L}\ \sum_{i=1}^{d}\lambda_i\chi^2_1, \tag{51}$$

*where* $\lambda_i$*'s are eigenvalues of* $\boldsymbol{\Sigma}^{1/2}\,\nabla_{\boldsymbol{\theta}}g(\boldsymbol{\theta}_0)\,\boldsymbol{\Sigma}^{1/2}$*. Moreover, variance of this asymptotic distribution can be written as*

$$\frac{1}{2}\left\|\boldsymbol{\Sigma}^{1/2}\,\nabla^2_{\mathbf{x}}g(\mathbf{x}_0)\,\boldsymbol{\Sigma}^{1/2}\right\|^2_F, \tag{52}$$

*where* $\|\cdot\|_F$ *is the Frobenius norm.*

**Proof** For proof see Appendix B. ■

### A.2 Parameter Estimation

Now suppose that the set of independent and identically distributed (i.i.d) set of samples $\mathbf{x}_1,...\,\mathbf{x}_n$ are generated from an underlying distribution that belongs to a parametric family, for which the density function $p(\mathbf{x}\,|\,\boldsymbol{\theta})$ can be represented by a multivariate parameter vector $\boldsymbol{\theta}$. Assume the true parameter is $\boldsymbol{\theta}_0$, that is $\{\mathbf{x}_i\}\sim p(\mathbf{x}\,|\,\boldsymbol{\theta}_0), i=1,...,n$. An *estimator* $\boldsymbol{\theta}_n=\boldsymbol{\theta}(\mathbf{x}_1,...,\mathbf{x}_n)$ is a function that maps the observed random variables to a point in the parameter space $\Omega$. The subscript $n$ in $\boldsymbol{\theta}_n$ indicates its dependence on the sample size. Since the observations are generated randomly, the estimators are also random and thus $\{\boldsymbol{\theta}_n\}$ 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 18 (Consistency)** *We say that an estimator* $\boldsymbol{\theta}_n$ *is* consistent *iff* $\boldsymbol{\theta}_n\xrightarrow{P}\boldsymbol{\theta}_0$*.*

Based on Theorem 14, 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 19 (Cramér-Rao, Lehmann and Casella 1998)** *Let* $\mathbf{x}_1,...,\mathbf{x}_n\sim p(\mathbf{x}\,|\,\boldsymbol{\theta}_0)$ *and* $\boldsymbol{\theta}_n=\boldsymbol{\theta}(\mathbf{x}_1,...,\mathbf{x}_n)$ *be an estimator. If the first moment of* $\boldsymbol{\theta}_n$ *is differentiable with respect to the parameter vector and its second moment is finite, then the following inequality holds for every* $\boldsymbol{\theta}\in\Omega$*:*

$$\mathrm{Cov}[\boldsymbol{\theta}_n]\ \succeq\ -\left(\nabla_{\boldsymbol{\theta}}\mathbb{E}[\boldsymbol{\theta}_n]\right)^{\top}\mathbf{I}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}\mathbb{E}[\boldsymbol{\theta}_n]. \tag{53}$$

The right-hand-side of (53) is called the *Cramér-Rao bound* of the estimator, where the middle term is the inverse of the *Fisher information matrix* of the parametric distribution $p(\mathbf{x}\,|\,\boldsymbol{\theta})$, defined as

$$\mathbf{I}(\boldsymbol{\theta})\ :=\ \mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\log p(\mathbf{x}\,|\,\boldsymbol{\theta})\cdot\nabla^{\top}_{\boldsymbol{\theta}}\log p(\mathbf{x}\,|\,\boldsymbol{\theta})\right].$$

Theorem 19 suggests that for an unbiased estimator $\boldsymbol{\theta}_n$, the inequality over the covariance matrix becomes: $\mathrm{Cov}[\boldsymbol{\theta}_n] \succeq \mathbf{I}(\boldsymbol{\theta})^{-1}, \forall\, \boldsymbol{\theta} \in \Omega$.

**Definition 20 (Efficiency)** *We say that an estimator $\boldsymbol{\theta}_n$ is* efficient*, iff it attains the Cramér-Rao bound, meaning that* $\mathrm{Cov}[\boldsymbol{\theta}_n]$ *achieves the lower-bound in (53) for every $n = 1, 2, \ldots$ . Furthermore, we say that $\boldsymbol{\theta}_n$ is* asymptotically efficient*, iff the lower bound is attained asymptotically (when $n \to \infty$).*

## Appendix 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 10) and the latter implies *boundedness in probability*. In what follows, if otherwise stated, $\{\boldsymbol{\theta}_n\}$ 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 (1991, Chapter 6).

**Definition 21** *We write $\boldsymbol{\theta}_n = o_p(a_n)$ iff*

$$\frac{\theta_{in}}{a_n} \;=\; o_p(1), \quad \textit{for all } i = 1, \ldots, d \tag{54}$$

**Definition 22** *We write $\boldsymbol{\theta}_n = O_p(a_n)$ iff the sequence $\left\{\frac{\theta_{in}}{a_n}\right\}$ is* bounded in probability *for every $i = 1, \ldots, d$, that is, for every $\epsilon > 0$ there exists $\delta_\epsilon$ such that*

$$P\left(\left|\frac{\theta_{in}}{a_n}\right| > \delta_\epsilon\right) \;<\; \epsilon, \quad n = 1, 2, \ldots \tag{55}$$

We also need the following propositions:

**Proposition 23 (Brockwell and Davis, 1991)** *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 24** *The followings are true:*[9]

*(i)* $\boldsymbol{\theta}_n = o_p(a_n) \;\Leftrightarrow\; \| \boldsymbol{\theta}_n \| = o_p(a_n)$.

*(ii)* $\boldsymbol{\theta}_n = O_p(a_n) \;\Leftrightarrow\; \| \boldsymbol{\theta}_n \| = O_p(a_n)$.

9. Unless subscripted otherwise, $\| \cdot \|$ denotes the $L_2$ norm in all the equations.

**Proof** The proof of part (i) can be found in Brockwell and Davis (1991, Proposition 6.1.2). Here, we only prove part (ii):

$(ii, \Rightarrow)$ : Since $\boldsymbol{\theta}_n = O_p(a_n)$, for every $\varepsilon > 0$ and for every $i = 1, ..., d$, there exists a coefficient $\delta_i > 0$ such that

$$P\left(|\theta_{ni}| > a_n \cdot \delta_i\right) \ < \ \frac{\varepsilon}{d} \quad , n = 1, 2, .... \tag{56}$$

Define $\delta_{\max} = \max\{\delta_1, ..., \delta_d\}$ and note that we can write

$$\begin{aligned} \left\{\boldsymbol{\theta}_n : \ \sum_{i=1}^{d} |\theta_{ni}|^2 > \left(d \cdot a_n \cdot \delta_{\max}\right)^2\right\} \ &\subseteq \ \left[\bigcap_{i=1}^{d} \{\boldsymbol{\theta}_n : \ |\theta_{ni}| \leq a_n \cdot \delta_{\max}\}\right]^c \\ &= \ \bigcup_{i=1}^{d} \{\boldsymbol{\theta}_n : \ |\theta_{ni}| > a_n \cdot \delta_{\max}\}, \end{aligned} \tag{57}$$

implying that

$$\begin{aligned} P\left(\| \boldsymbol{\theta}_n \|^2 > \left(d \cdot a_n \cdot \delta_{\max}\right)^2\right) \ &\leq \ P\left(\bigcup_{i=1}^{d} \{\boldsymbol{\theta}_n : \ |\theta_{ni}| > a_n \cdot \delta_{\max}\}\right) \\ &\leq \ \sum_{i=1}^{d} P(|\theta_{ni}| > a_n \cdot \delta_{\max}). \end{aligned} \tag{58}$$

Furthermore, for every $i \in \{1, ..., d\}$, 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

$$P\left(\| \boldsymbol{\theta}_n \|^2 > \left(d \cdot a_n \cdot \delta_{\max}\right)^2\right) \ \leq \ \sum_{i=1}^{d} P(|\theta_{ni}| > a_n \cdot \delta_i) \ < \ \varepsilon. \tag{59}$$

Therefore, for every $\varepsilon > 0$, we can choose $\delta_\varepsilon = d \cdot \delta_{\max}$ such that $P\left(\frac{\| \boldsymbol{\theta}_n \|}{a_n} > \delta_\varepsilon\right) < \varepsilon$ for every $n = 1, 2, ....$ Therefore, by definition, $\| \boldsymbol{\theta}_n \| = O_p(a_n)$.
$(ii, \Leftarrow)$ : Suppose $\| \boldsymbol{\theta}_n \| = O_p(a_n)$, that is for every $\varepsilon > 0$ we can find $\delta_\varepsilon > 0$ such that

$$P(\| \boldsymbol{\theta}_n \| > a_n \cdot \delta_\varepsilon) \ < \varepsilon \quad , n = 1, 2, ... \tag{60}$$

It is clear that for any given $i \in \{1, ..., d\}$ we have

$$\{\boldsymbol{\theta}_n : |\theta_{ni}| > a_n \cdot \delta_\varepsilon\} \ \subseteq \ \{\boldsymbol{\theta}_n : \| \boldsymbol{\theta}_n \| > a_n \cdot \delta_\varepsilon\}, \tag{61}$$

hence

$$P(|\theta_{ni}| > a_n \cdot \delta_\varepsilon) \ \leq \ P(\| \boldsymbol{\theta}_n \| > a_n \cdot \delta_\varepsilon) \ < \epsilon \quad , n = 1, 2, ... \tag{62}$$

meaning that $\theta_{ni} = O_p(a_n), \forall i \in \{1, ..., d\}$, or equivalently $\boldsymbol{\theta}_n = O_p(a_n)$. ■

**Proposition 25** *If $\boldsymbol{\theta}_n = O_p(a_n)$ and $a_n \to 0 (as\ n \to \infty)$, then $\boldsymbol{\theta}_n = o_p(1)$.*

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

$$P(\| \boldsymbol{\theta}_n \| > \varepsilon) < \varepsilon_0 \quad \forall n > N. \tag{63}$$

Because of the assumption of being bounded by $a_n$, that is $\boldsymbol{\theta}_n = O_p(a_n)$ or equivalently $\| \boldsymbol{\theta}_n \| = O_p(a_n)$, we can choose a real number $\delta_0 > 0$ such that

$$P(\| \boldsymbol{\theta}_n \| > a_n \delta_0) \ < \ \varepsilon_0 \quad n = 1, 2, ... \tag{64}$$

On the other hand, since $a_n \to 0 (\text{as } n \to \infty)$, 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

$$[0, a_n \delta_0] \ \subseteq \ [0, \varepsilon] \quad \forall n > N_0, \tag{65}$$

implying that

$$P(\| \boldsymbol{\theta}_n \| \leq a_n \delta_0) \ \leq \ P(\| \boldsymbol{\theta}_n \| \leq \varepsilon) \quad \forall n > N_0. \tag{66}$$

From inequalities (64) and (66), and noticing that the latter holds for all $n$ whereas the former is satisfied when $n > N_0$, one can write:

$$P(\| \boldsymbol{\theta}_n \| > \varepsilon) \ \leq \ P(\| \boldsymbol{\theta}_n \| > a_n \delta_0) \ < \ \varepsilon_0 \quad \forall n > N_0. \tag{67}$$

Therefore, for every $\varepsilon_0 > 0$, equation (63) is guaranteed if $N$ is chosen to be equal to $N_0$ so that inequality (66) is satisfied. Similarly, this can be written for every $\varepsilon > 0$, thus the proof is complete. ■

**Proposition 26 (Serfling 2002, Chapter 1)** *Let $\{\boldsymbol{\theta}_n\}$ be a sequence of random variables. If there exists a random variable $\tilde{\boldsymbol{\theta}}$ such that $\boldsymbol{\theta}_n \xrightarrow{L} \tilde{\boldsymbol{\theta}}$, then $\boldsymbol{\theta}_n = O_p(1)$.*

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

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

**Theorem 27** *Let $\{\boldsymbol{\theta}_n\}$ be a sequence of random vectors in a convex and compact set $\Omega \subseteq \mathbb{R}^d$ and $\boldsymbol{\theta}_0 \in \Omega$ be a constant vector such that $\boldsymbol{\theta}_n - \boldsymbol{\theta}_0 = O_p(a_n)$ where $a_n \to 0 (\text{as } n \to \infty)$. If $g : \Omega \to \mathbb{R}$ is a $\mathcal{C}^3$ function , then*

$$g(\boldsymbol{\theta}_n) \ = \ g(\boldsymbol{\theta}_0) \ + \ \nabla_{\boldsymbol{\theta}}^{\top} g(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) \ + \ \frac{1}{2}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)^{\top} \nabla_{\boldsymbol{\theta}}^2 g(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) \ + \ o_p(a_n^2). \tag{68}$$

**Proof** Since $g$ is twice continuously differentiable in a neighborhood of $\boldsymbol{\theta}_0$, it can be written in terms of the Taylor expansion as

$$g(\boldsymbol{\theta}) \ = \ g(\boldsymbol{\theta}_0) \ + \ (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\top} \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_0) \ + \ \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\top} \nabla_{\boldsymbol{\theta}}^2 g(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \ + \ r_2(\boldsymbol{\theta}, \boldsymbol{\theta}_0), \tag{69}$$

where $r_2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ 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 $\boldsymbol{\theta}^* = t\,\boldsymbol{\theta} + (1-t)\,\boldsymbol{\theta}_0 \in \Omega$ (due to convexity of $\Omega$) and

$$r_2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \;=\; \frac{1}{6} \sum_{1 \le i,j,k \le d} \frac{\partial^3 g(\boldsymbol{\theta}^*)}{\partial\theta_i \partial\theta_j \partial\theta_k} (\theta_i - \theta_{0i})(\theta_j - \theta_{0j})(\theta_k - \theta_{0k}). \tag{70}$$

But since $\Omega$ is compact and $g \in \mathcal{C}^3$, the third derivative of $g$ is bounded[10] and therefore there exists $M > 0$ such that

$$\left| \frac{\partial^3 g(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j \partial\theta_k} \right| \;\le\; M \quad , \forall\, \boldsymbol{\theta} \in \Omega \;\; , \forall i,j,k \in \{1, ..., d\}. \tag{71}$$

Hence, the Lagrange remainder can be bounded by

$$\begin{aligned} |r_2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)| \;&\le\; \frac{M}{6} \sum_{1 \le i,j,k \le 3} |\theta_i - \theta_{0i}| \cdot |\theta_j - \theta_{0j}| \cdot |\theta_k - \theta_{0k}| \\ &=\; \frac{M}{6} \| \boldsymbol{\theta} - \boldsymbol{\theta}_0 \|_1^3 \\ &\le\; \frac{M'}{6} \| \boldsymbol{\theta} - \boldsymbol{\theta}_0 \|^3, \end{aligned} \tag{72}$$

where $M' = c_u M$ and $c_u$ is obtained from the equivalence of norms in $\mathbb{R}^d$ vector space.[11] Now define the function $h : \Omega \to \mathbb{R}$ as below

$$h(\boldsymbol{\theta}) \;:=\; \begin{cases} \dfrac{r_2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)}{\| \boldsymbol{\theta} - \boldsymbol{\theta}_0 \|^2 / 2} & , \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \\ 0 & , \boldsymbol{\theta} = \boldsymbol{\theta}_0 \end{cases} \tag{74}$$

Note that $h(\boldsymbol{\theta})$ is continuous at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$: due to boundedness of $r_2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$, $h(\boldsymbol{\theta})$ is also bounded by

$$|h(\boldsymbol{\theta})| \;\le\; \frac{M'}{3} \| \boldsymbol{\theta} - \boldsymbol{\theta}_0 \|. \tag{75}$$

Hence, for every $\varepsilon > 0$, we can select $\delta_\varepsilon = \frac{3\varepsilon}{M'}$ such that the following continuity condition holds:

$$\| \boldsymbol{\theta} - \boldsymbol{\theta}_0 \| < \delta_\varepsilon \;\; \Rightarrow \;\; |h(\boldsymbol{\theta})| \;\le\; \varepsilon. \tag{76}$$

10. This is because of the following Theorem in real analysis:

    **Theorem 28** *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, where $Y = \mathbb{R}$, compactness of $f(X)$ is equivalent to boundedness and closedness.

11. Two norm functions $\| \cdot \|_{(1)}$ and $\| \cdot \|_{(2)}$ in a vector space $\Omega$, are called *equivalent* iff there exist constants $c_u \ge c_d > 0$ such that

$$c_d \| \boldsymbol{\theta} \|_{(2)} \;\le\; \| \boldsymbol{\theta} \|_{(1)} \;\le\; c_u \| \boldsymbol{\theta} \|_{(2)} \quad , \forall\, \boldsymbol{\theta} \in \Omega. \tag{73}$$

Continuity of $h(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ implies $\lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}_0} h(\boldsymbol{\theta}) = h(\boldsymbol{\theta}_0) = 0$. Furthermore, since $\boldsymbol{\theta}_n - \boldsymbol{\theta}_0 = O_p(a_n)$ and $a_n \to 0$(as $n \to \infty$), Proposition 25 suggests that $\boldsymbol{\theta}_n - \boldsymbol{\theta}_0 = o_p(1)$. These two enable us to use Proposition 11 and write

$$h(\boldsymbol{\theta}_n) - h(\boldsymbol{\theta}_0) \;=\; h(\boldsymbol{\theta}_n) \;=\; o_p(1). \tag{77}$$

Finally, from equation (74) and Propositions 23, 24 and 25, we can write that

$$r_2(\boldsymbol{\theta}_n, \boldsymbol{\theta}_0) \;=\; h(\boldsymbol{\theta}_n) \cdot \frac{\| \boldsymbol{\theta}_n - \boldsymbol{\theta}_0 \|^2}{2} \;=\; o_p(1) \cdot O_p(a_n^2) \;=\; o_p(a_n^2) \tag{78}$$

■

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

Finally, here is the proof of second-order multivariate Delta method (Theorem 17):

**Proof** From the assumption for convergence in law, i.e., $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, and Proposition 26, one concludes that $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) = O_p(1)$ and therefore $\boldsymbol{\theta}_n - \boldsymbol{\theta}_0 = O_p\left(\frac{1}{\sqrt{n}}\right)$. Thus, we can use Theorem 27 with $a_n = \frac{1}{\sqrt{n}}$ to write

$$g(\boldsymbol{\theta}) \;=\; g(\boldsymbol{\theta}_0) \;+\; (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_0) \;+\; \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \nabla^2_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \;+\; o_p\left(\frac{1}{n}\right), \tag{79}$$

hence

$$\begin{aligned} n\Big[g(\boldsymbol{\theta}) - g(\boldsymbol{\theta}_0)\Big] \;&=\; \frac{1}{2}\left[\sqrt{n}\cdot(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right]^\top \nabla^2_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_0) \left[\sqrt{n}\cdot(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right] \;+\; o_p(1) \\ &\xrightarrow{L}\; \frac{1}{2}\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})^\top \nabla^2_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_0)\, \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \\ &=\; \frac{1}{2}\mathcal{N}(\mathbf{0}, \mathbb{I}_d)^\top \left[\boldsymbol{\Sigma}^{1/2}\, \nabla^2_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_0)\, \boldsymbol{\Sigma}^{1/2}\right] \mathcal{N}(\mathbf{0}, \mathbb{I}_d). \end{aligned} \tag{80}$$

Define $\boldsymbol{\Gamma} := \boldsymbol{\Sigma}^{1/2}\, \nabla^2_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_0)\, \boldsymbol{\Sigma}^{1/2}$ and rewrite the right-hand-side element-wise as

$$\frac{1}{2}\mathcal{N}(\mathbf{0}, \mathbb{I}_d)^\top \boldsymbol{\Gamma}\, \mathcal{N}(\mathbf{0}, \mathbb{I}_d) \;=\; \frac{1}{2}\sum_{i=1}^{d} \lambda_i\, \mathcal{N}(0,1)^2 \;=\; \frac{1}{2}\sum_{i=1}^{d} \lambda_i \chi^2_1, \tag{81}$$

where $\lambda_i$'s are eigenvalues of $\boldsymbol{\Gamma}$. Finally, noting that the terms in the Chi-square mixture are independent, variance of the convergent random variable can be easily computed as

$$\begin{aligned} \mathrm{Var}\left[\frac{1}{2}\sum_{i=1}^{d} \lambda_i \chi^2_1\right] \;&=\; \frac{1}{4}\sum_{i=1}^{d} \lambda_i^2 \cdot \mathrm{Var}\left[\chi^2_1\right] \\ &=\; \frac{1}{2}\sum_{i=1}^{d} \lambda_i^2 \\ &=\; \frac{1}{2}\left\| \boldsymbol{\Sigma}^{1/2}\, \nabla^2_{\mathbf{x}} g(\mathbf{x}_0)\, \boldsymbol{\Sigma}^{1/2} \right\|_F^2. \end{aligned} \tag{82}$$

■

## Appendix C. Proof of Lemma 7

We first substitute the score function of the classifier, i.e.,

$$\nabla_{\boldsymbol{\theta}} \log p(y|\,\mathbf{x},\boldsymbol{\theta}) \;=\; \frac{\nabla_{\boldsymbol{\theta}} p(y|\,\mathbf{x},\boldsymbol{\theta})}{p(y|\,\mathbf{x},\boldsymbol{\theta})},$$

into Monte-Carlo approximation of $\mathbf{I}_q$ to get

$$\begin{aligned}\hat{\mathbf{I}}(\boldsymbol{\theta};X_q) \;&=\; \frac{1}{|X_q|}\sum_{\mathbf{x}\in X_q}\sum_{y=1}^{c} p(y|\,\mathbf{x},\boldsymbol{\theta})\cdot\frac{\nabla_{\boldsymbol{\theta}} p(y|\,\mathbf{x},\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}^{\top} p(y|\,\mathbf{x},\boldsymbol{\theta})}{p(y|\,\mathbf{x},\boldsymbol{\theta})^2} \;+\; \delta\mathbb{I}_d \\ &=\; \frac{1}{|X_q|}\sum_{\mathbf{x}\in X_q}\sum_{y=1}^{c}\frac{\nabla_{\boldsymbol{\theta}} p(y|\,\mathbf{x},\boldsymbol{\theta}).\nabla_{\boldsymbol{\theta}}^{\top} p(y|\,\mathbf{x},\boldsymbol{\theta})}{p(y|\,\mathbf{x},\boldsymbol{\theta})} \;+\; \delta\mathbb{I}_d. \end{aligned} \tag{83}$$

Define the vector $\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y) := \nabla_{\boldsymbol{\theta}} p(y|\,\mathbf{x},\boldsymbol{\theta})\Big/\sqrt{p(y|\,\mathbf{x},\boldsymbol{\theta})}$ and rewrite $\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)$ as

$$\hat{\mathbf{I}}(\boldsymbol{\theta};X_q) \;=\; \frac{1}{|X_q|}\sum_{\mathbf{x}\in X_q}\sum_{y=1}^{c}\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y).\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)^{\top} + \delta\cdot\mathbb{I}_d. \tag{84}$$

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

$$\begin{aligned}\hat{\mathbf{I}}(\boldsymbol{\theta};X_p) \;&=\; \frac{|X_q|}{|X_p|}\left[\frac{1}{|X_q|}\sum_{\mathbf{x}\in X_q}\sum_{y=1}^{c}\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y).\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)^{\top} + \delta\cdot\mathbb{I}_d\right] \\ &\quad+\; \frac{1}{|X_p|}\sum_{\mathbf{x}\in X_p - X_q}\sum_{y=1}^{c}\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y).\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)^{\top} \;+\; \delta\left(\frac{|X_p|-|X_q|}{|X_p|}\right)\cdot\mathbb{I}_d \\ &=\; \left(\frac{|X_q|}{|X_p|}\right)\cdot\hat{\mathbf{I}}(\boldsymbol{\theta};X_q) \;+\; \frac{1}{|X_p|}\sum_{\mathbf{x}\in X_p - X_q}\sum_{y=1}^{c}\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y).\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)^{\top} \\ &\quad+\; \delta\left(\frac{|X_p|-|X_q|}{|X_p|}\right)\cdot\mathbb{I}_d \end{aligned} \tag{85}$$

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

$$\begin{aligned}\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)^{-1}\hat{\mathbf{I}}(\boldsymbol{\theta};X_p) \;&=\; \left(\frac{|X_q|}{|X_p|}\right)\cdot\mathbb{I}_d \;+\; \frac{\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)^{-1}}{|X_p|}\left[\sum_{\mathbf{x}\in X_p - X_q}\sum_{y=1}^{c}\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\cdot\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)^{\top}\right] \\ &\quad+\; \delta\left(\frac{|X_p|-|X_q|}{|X_p|}\right)\cdot\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)^{-1}. \end{aligned} \tag{86}$$

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

$$\begin{aligned}\operatorname{tr}\left[\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)^{-1}\hat{\mathbf{I}}(\boldsymbol{\theta};X_p)\right] &= \frac{|X_q|\cdot d}{|X_p|}+\frac{1}{|X_p|}\sum_{\mathbf{x}\in X_p-X_q}\sum_{y=1}^{c}\operatorname{tr}\left[\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)^{-1}\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\cdot\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)^{\top}\right]\\ &\quad+\ \delta\left(\frac{|X_p|-|X_q|}{|X_p|}\right)\cdot\operatorname{tr}\left[\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)^{-1}\right]\\ &\approx \frac{|X_q|\cdot d}{|X_p|}\ +\ \frac{1}{|X_p|}\sum_{\mathbf{x}\in X_p-X_q}\sum_{y=1}^{c}\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)^{\top}\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)^{-1}\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y),\end{aligned} \tag{87}$$

where the last term is dropped since the overloading constant, $\delta$, is assumed to be small. Furthermore, the term including $\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)^{-1}$ can be approximated by replacing the weighted harmonic mean of the eigenvalues of $\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)$ by their weighted arithmetic mean (Hoi et al., 2006)

$$\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)^{\top}\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)^{-1}\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\ \approx\ \frac{\|\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^4}{\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)^{\top}\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)}. \tag{88}$$

Note that this approximation becomes exact when the condition number of $\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)$ is one. Substituting $\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)$ from equation (84) into the denominator of the approximation above yields

$$\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)^{\top}\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\ =\ \frac{1}{|X_q|}\sum_{\mathbf{x}'\in X_q}\sum_{y'=1}^{c}\left[\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)^{\top}\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}',y')\right]^2\ +\ \delta\|\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2 \tag{89}$$

Assume that the value of $\boldsymbol{\theta}$ is not located at the stationary point of the conditional density $p(y|\,\mathbf{x},\boldsymbol{\theta})$, hence $\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)$ is not the zero vector. Integrating approximation (88) with equation (87) results

$$\begin{aligned}\operatorname{tr}\left[\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)^{-1}\hat{\mathbf{I}}(\boldsymbol{\theta};X_p)\right] &\approx \frac{|X_q|\cdot d}{|X_p|}\\ &\quad+\ \frac{1}{|X_p|}\sum_{\mathbf{x}\in X_p-X_q}\sum_{y=1}^{c}\frac{1}{\delta\cdot\|\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^{-2}+\sum_{\mathbf{x}'\in X_q}g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}')},\end{aligned} \tag{90}$$

where

$$g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}')\ :=\ \frac{1}{|X_q|}\sum_{y'=1}^{c}\left[\frac{\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)^{\top}\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}',y')}{\|\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2}\right]^2. \tag{91}$$

Finally by removing the constants in (90), we get the optimization

$$\begin{aligned}&\underset{X_q\subset X_p}{\arg\min}\ \operatorname{tr}\left[\hat{\mathbf{I}}(\boldsymbol{\theta};X_q)^{-1}\hat{\mathbf{I}}(\boldsymbol{\theta};X_p)\right]\\ &\qquad\approx\ \underset{X_q\subset X_p}{\arg\max}\sum_{\mathbf{x}\in X_p-X_q}\sum_{y=1}^{c}\frac{-1}{\delta\cdot\|\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^{-2}+\sum_{\mathbf{x}'\in X_q}g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}')}.\end{aligned} \tag{92}$$

## Appendix D. Proof of Theorem 8

Proof of this Theorem is a generalization of the discussion by Hoi et al. (2006), with clarification of all the assumptions and approximations made.

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

$$f_{\boldsymbol{\theta}}(X_q; y) \;=\; \sum_{\mathbf{x} \in X_p - X_q} \frac{-1}{\delta \cdot \|v_{\boldsymbol{\theta}}(\mathbf{x}, y)\|^{-2} + \sum_{\mathbf{x}' \in X_q} g_{\boldsymbol{\theta}}(\mathbf{x}, y, \mathbf{x}')}, \quad \forall X_q \subseteq X_p. \tag{93}$$

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

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

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

$$\begin{aligned} \rho_{f_{\boldsymbol{\theta}}(\cdot;y)}(X_q; \boldsymbol{\xi}) \;&=\; f_{\boldsymbol{\theta}}(X_q \cup \{\boldsymbol{\xi}\}; y) \;-\; f_{\boldsymbol{\theta}}(X_q; y) \\ &= \sum_{\mathbf{x} \in X_p - X_q} \left[ \frac{-1}{\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}, y)\|^2} + \sum_{\mathbf{x}' \in X_q \cup \{\boldsymbol{\xi}\}} g_{\boldsymbol{\theta}}(\mathbf{x}, y, \mathbf{x}')} + \frac{1}{\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}, y)\|^2} + \sum_{\mathbf{x}' \in X_q} g_{\boldsymbol{\theta}}(\mathbf{x}, y, \mathbf{x}')} \right] \\ &\quad + \frac{1}{\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}, y)\|^2} + \sum_{\mathbf{x}' \in X_q \cup \{\boldsymbol{\xi}\}} g_{\boldsymbol{\theta}}(\boldsymbol{\xi}, y, \mathbf{x}')}. \end{aligned}$$

The right-hand-side can be rewritten as

$$\begin{aligned} &\sum_{\mathbf{x} \in X_p - X_q} \left[ \frac{g_{\boldsymbol{\theta}}(\mathbf{x}, y, \boldsymbol{\xi})}{\left( \frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}, y)\|^2} + \sum_{\mathbf{x}' \in X_q \cup \{\boldsymbol{\xi}\}} g_{\boldsymbol{\theta}}(\mathbf{x}, y, \mathbf{x}') \right) \left( \frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}, y)\|^2} + \sum_{\mathbf{x}' \in X_q} g_{\boldsymbol{\theta}}(\mathbf{x}, y, \mathbf{x}') \right)} \right] \\ &\quad + \frac{1}{\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}, y)\|^2} + \sum_{\mathbf{x}' \in X_q \cup \{\boldsymbol{\xi}\}} g_{\boldsymbol{\theta}}(\boldsymbol{\xi}, y, \mathbf{x}')}. \end{aligned} \tag{94}$$

Since by definition $g_{\boldsymbol{\theta}}(\mathbf{x}, y, \mathbf{x}') \geq 0, \forall\, \mathbf{x}, y, \mathbf{x}'$, all of the terms in (94) are non-negative and therefore $\rho_{f_{\boldsymbol{\theta}}(\cdot;y)}(X_q; \boldsymbol{\xi}) \geq 0$. This is true for any $X_q \subseteq X_p$, hence monotonicity of $f_{\boldsymbol{\theta}}(\cdot; y)$ is obtained. Now let us take any superset $X_{q'}$ such that $X_q \subseteq X_{q'} \subseteq X_p$ and $\boldsymbol{\xi} \in X_p - X_{q'}$, and form the difference between their corresponding discrete derivatives. From (94), we will

have

$$\begin{aligned}
&\rho_{f_{\boldsymbol{\theta}}(\cdot;y)}(X_q;\boldsymbol{\xi}) - \rho_{f_{\boldsymbol{\theta}}(\cdot;y)}(X_{q'};\boldsymbol{\xi}) \\
&= \sum_{\mathbf{x}\in X_p - X_q} \left[ \frac{g_{\boldsymbol{\theta}}(\mathbf{x},y,\boldsymbol{\xi})}{\left(\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2} + \sum_{\mathbf{x}'\in X_q\cup\{\boldsymbol{\xi}\}} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}')\right)\left(\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2} + \sum_{\mathbf{x}'\in X_q} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}')\right)} \right] \\
&+ \frac{1}{\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\boldsymbol{\xi},y)\|^2} + \sum_{\mathbf{x}'\in X_q\cup\{\boldsymbol{\xi}\}} g_{\boldsymbol{\theta}}(\boldsymbol{\xi},y,\mathbf{x}')} \\
&- \sum_{\mathbf{x}\in X_p - X_{q'}} \left[ \frac{g_{\boldsymbol{\theta}}(\mathbf{x},y,\boldsymbol{\xi})}{\left(\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2} + \sum_{\mathbf{x}'\in X_{q'}\cup\{\boldsymbol{\xi}\}} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}')\right)\left(\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2} + \sum_{\mathbf{x}'\in X_{q'}} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}')\right)} \right] \\
&- \frac{1}{\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\boldsymbol{\xi},y)\|^2} + \sum_{\mathbf{x}'\in X_{q'}\cup\{\boldsymbol{\xi}\}} g_{\boldsymbol{\theta}}(\boldsymbol{\xi},y,\mathbf{x}')}.
\end{aligned} \tag{95}$$

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

$$\begin{aligned}
&\sum_{\mathbf{x}'\in X_{q'}} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}') \;\geq\; \sum_{\mathbf{x}'\in X_q} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}') \\
\Leftrightarrow\quad &\left[\sum_{\mathbf{x}'\in X_{q'}} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}') + \frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2}\right]^{-1} \leq \left[\sum_{\mathbf{x}'\in X_q} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}') + \frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2}\right]^{-1} \\
\Leftrightarrow\quad &-\left[\sum_{\mathbf{x}'\in X_{q'}} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}') + \frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2}\right]^{-1} \geq -\left[\sum_{\mathbf{x}'\in X_q} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}') + \frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2}\right]^{-1}
\end{aligned} \tag{96}$$

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

$$-\left[\sum_{\mathbf{x}'\in X_{q'}\cup\{\boldsymbol{\xi}\}} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}') + \frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2}\right]^{-1} \geq -\left[\sum_{\mathbf{x}'\in X_q\cup\{\boldsymbol{\xi}\}} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}') + \frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2}\right]^{-1} \tag{97}$$

Applying the inequalities (96) and (97) into euqation (95) results

$$\begin{aligned}
&\rho_{f_{\boldsymbol{\theta}}(\cdot;y)}(X_q;\boldsymbol{\xi}) - \rho_{f_{\boldsymbol{\theta}}(\cdot;y)}(X_{q'};\boldsymbol{\xi}) \\
&\geq \sum_{\mathbf{x}\in X_p - X_q} \left[ \frac{g_{\boldsymbol{\theta}}(\mathbf{x},y,\boldsymbol{\xi})}{\left(\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2} + \sum_{\mathbf{x}'\in X_q\cup\{\boldsymbol{\xi}\}} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}')\right)\left(\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2} + \sum_{\mathbf{x}'\in X_q} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}')\right)} \right] \\
&- \sum_{\mathbf{x}\in X_p - X_{q'}} \left[ \frac{g_{\boldsymbol{\theta}}(\mathbf{x},y,\boldsymbol{\xi})}{\left(\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2} + \sum_{\mathbf{x}'\in X_q\cup\{\boldsymbol{\xi}\}} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}')\right)\left(\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2} + \sum_{\mathbf{x}'\in X_q} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}')\right)} \right] \\
&+ \frac{1}{\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\boldsymbol{\xi},y)\|^2} + \sum_{\mathbf{x}'\in X_q\cup\{\boldsymbol{\xi}\}} g_{\boldsymbol{\theta}}(\boldsymbol{\xi},y,\mathbf{x}')} - \frac{1}{\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\boldsymbol{\xi},y)\|^2} + \sum_{\mathbf{x}'\in X_q\cup\{\boldsymbol{\xi}\}} g_{\boldsymbol{\theta}}(\boldsymbol{\xi},y,\mathbf{x}')},
\end{aligned} \tag{98}$$

which yields[12]

$$\sum_{\mathbf{x}\in X_{q'}-X_q} \frac{g_{\boldsymbol{\theta}}(\mathbf{x},y,\boldsymbol{\xi})}{\left(\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2}+\sum_{\mathbf{x}'\in X_q\cup\{\boldsymbol{\xi}\}} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}')\right)\left(\frac{\delta}{\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x},y)\|^2}+\sum_{\mathbf{x}'\in X_q} g_{\boldsymbol{\theta}}(\mathbf{x},y,\mathbf{x}')\right)} \geq 0. \tag{99}$$

Inequality (99) holds for any $X_q \subseteq X_p$; hence submodularity of $f_{\boldsymbol{\theta}}(\cdot;y)$ stands for all $y \in \{1,...,c\}$ and $\boldsymbol{\theta} \in \Omega$.

## References


Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. *arXiv preprint arXiv:1307.8371*, 2013.

Javad Azimi, Alan Fern, Xiaoli Zhang-Fern, Glencora Borradaile, and Brent Heeringa. Batch active learning via coordinated matching. *arXiv preprint arXiv:1206.6458*, 2012.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 65–72. ACM, 2006.

Alina Beygelzimer, John Langford, Zhang Tong, and Daniel J Hsu. Agnostic active learning without constraints. In *Advances in Neural Information Processing Systems*, pages 199–207, 2010.

Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 1. Springer New York, 2006.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Peter J Brockwell and Richard A Davis. *Time Series: Theory and Methods*. Springer, 1991.

Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. *arXiv preprint arXiv:1506.02348*, 2015a.

Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. In *Advances in Neural Information Processing Systems*, pages 1090–1098, 2015b.

Yuxin Chen and Andreas 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.


12. The inequality in (98) is obtained by the fact that, for every four positive real numbers $a$, $a_0$, $b$ and $b_0$, if we have $-a \geq -a_0$ and $-b \geq -b_0$ (similar to equations 96 and 97), then

$$-a\cdot b \;=\; (-a)\cdot b \;\geq\; (-a_0)\cdot b \;=\; a_0\cdot(-b) \;\geq\; a_0\cdot(-b_0) \;=\; -a_0\cdot b_0.$$

David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

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

David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *arXiv preprint cs/9603104*, 1996.

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

Sanjoy Dasgupta, Adam Tauman Kalai, and Claire Monteleoni. Analysis of perceptron-based active learning. In *Learning Theory*, pages 249–263. Springer, 2005.

Sanjoy Dasgupta, Claire Monteleoni, and Daniel J Hsu. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems*, pages 353–360, 2007.

Richard O Duda, Peter E Hart, and David G Stork. *Pattern Classification*. John Wiley & Sons,, 1999.

Valerii Vadimovich Fedorov. *Theory of Optimal Experiments*. Elsevier, 1972.

Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

Yifan Fu, Xingquan Zhu, and Bin Li. A survey on instance selection for active learning. *Knowledge and information systems*, 35(2):249–283, 2013.

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

Yuhong Guo. Active instance sampling via matrix partition. In *Advances in Neural Information Processing Systems*, pages 802–810, 2010.

Yuhong Guo and Russell Greiner. Optimistic active-learning using mutual information. In *National Joint Conferences on Artificial Intelligence*, volume 7, pages 823–829, 2007.

Yuhong Guo and Dale Schuurmans. Discriminative batch mode active learning. In *Advances in Neural Information Processing Systems*, pages 593–600, 2008.

Steve Hanneke. Activized learning: Transforming passive to active with improved label complexity. *Journal of Machine Learning Research*, 13(May):1469–1587, 2012.

Steve Hanneke et al. Rates of convergence in active learning. *The Annals of Statistics*, 39 (1):333–361, 2011.

Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael 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.

Steven CH Hoi, Rong Jin, and Michael 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.

Alex Holub, Pietro Perona, and Michael 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.

Ming Ji and Jiawei Han. A variance minimization criterion to active learning on graphs. In *AISTATS*, pages 556–564, 2012.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *The Journal of Machine Learning Research*, 9:235–284, 2008.

Erich Leo Lehmann and George Casella. *Theory of Point Estimation*. Springer, 1998.

David MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

Noboru Murata, Shuji 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.

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

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functionsi. *Mathematical Programming*, 14(1): 265–294, 1978.

Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the 21st International Conference on Machine Learning*, page 79. ACM, 2004.

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

Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *International Conference on Machine Learning*, pages 839–846, 2000.

Robert J Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 2002.

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

Burr Settles and Mark 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.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in Neural Information Processing Systems*, pages 1289–1296, 2008.

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

Jamshid Sourati, Murat Akcakaya, Jennifer G Dy, Todd K Leen, and Deniz Erdogmus. Classification active learning based on mutual information. *Entropy*, 18(2):51, 2016.

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

Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM review*, 38(1): 49–95, 1996.

Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.

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

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

Tong 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.

Xiaojin Zhu, John Lafferty, and Zoubin 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.