MODEL CHANGE DETECTION WITH APPLICATION TO MACHINE LEARNING

Yuheng Bu[⋆] *Jiaxun Lu*[†] *Venugopal V. Veeravalli*[⋆]

[⋆] University of Illinois at Urbana-Champaign [†] Tsinghua University
Email: bu3@illinois.edu, lujx14@mails.tsinghua.edu.cn, vvv@illinois.edu

ABSTRACT

Model change detection is studied, in which there are two sets of samples that are independently and identically distributed (i.i.d.) according to a pre-change probabilistic model with parameter θ, and a post-change model with parameter θ', respectively. The goal is to detect whether the change in the model is significant, i.e., whether the difference between the pre-change parameter and the post-change parameter $\|\theta - \theta'\|_2$ is larger than a pre-determined threshold ρ. The problem is considered in a Neyman-Pearson setting, where the goal is to maximize the probability of detection under a false alarm constraint. Since the generalized likelihood ratio test (GLRT) is difficult to compute in this problem, we construct an empirical difference test (EDT), which approximates the GLRT and has low computational complexity. Moreover, we provide an approximation method to set the threshold of the EDT to meet the false alarm constraint. Experiments with linear regression and logistic regression are conducted to validate the proposed algorithms.

***Index Terms*—** Model change detection, generalized likelihood ratio test, Neyman-Pearson setting

1. INTRODUCTION

We study the model change detection problem, where two sets of samples are independently and identically distributed (i.i.d.) according to a pre-change probabilistic model with parameter θ, and a post-change probabilistic model with parameter θ', respectively. The goal is to determine whether the change in the model is significant or not. We formulate the problem in a Neyman-Pearson setting, and adopt the ℓ_2 distance between the parameters to measure the change between the models. More specifically, our goal is to construct a test to detect whether $\|\theta - \theta'\|_2$ is larger than a pre-determined threshold ρ, while satisfying a false alarm constraint.

This problem is motivated in part by the recent works on active and adaptive sequential learning [1–3], where the machine learning models learned in previous time-steps are used adaptively to improve the accuracy and data-efficiency in the next time-step. A key step in applying these adaptive sequential learning methods is the detection of an abrupt or large model change, since adapting to the previous model if it is significantly different from the current one could deteriorate performance. A specific application in this context is the detection of a shift of user preferences in personalized recommendation systems [4, 5]. In addition, we believe that our model change detection formulation can be applied in transfer learning [6] to determine whether two machine learning tasks are transferable.

We note that our model change detection problem is different from the quickest change detection problem studied in [7, 8]. There a linear regression model changes at an unknown point in time, and the goal is to detect the change as soon as possible with streaming data. We are interested in detecting whether the change in the model is significant, given sets of samples from the pre- and post-change models.

A standard method for solving a composite hypothesis testing problem such as the model change detection problem under consideration is the generalized likelihood ratio test (GLRT). However, the maximum likelihood estimates of θ and θ' required in the GLRT are difficult to compute under the constraint $\|\theta - \theta'\|_2 \le \rho$ in this case. Our first contribution is to propose an empirical difference test (EDT), which approximates the GLRT and has low computational complexity. Moreover, we provide an approximation method to set the threshold in the proposed EDT, which ensures a bound on the worst-case false alarm probability. We validate our results using experiments involving linear regression and logistic regression.

2. PROBLEM MODEL

Throughout this paper, we use lower case letters to denote scalars and vectors, and use upper case letters to denote random variables and matrices. We use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote the largest and the smallest eigenvalues of matrix A, respectively, and $\mathrm{Tr}(A)$ to denote the trace of a square matrix A. All logarithms are the natural ones.

We consider the model change detection problem in the following setting. We are given two datasets $\mathcal{S} = \{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$ and $\mathcal{S}' = \{\mathbf{z}'_1, \cdots, \mathbf{z}'_{n'}\}$ with samples $\mathbf{z}$ drawn from some instance space $\mathcal{Z}$. In addition, we are given a pa-

The work of Y. Bu and V. V. Veeravalli was supported by the Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196 through the University of Illinois at Urbana-Champaign.

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

rameterized family of distribution models $\mathcal{M} = \{p(\mathbf{z}|\theta), \theta \in \mathbb{R}^d\}$. We assume that there exist two unknown parameters $\theta, \theta' \in \mathbb{R}^d$, such that the datasets $\mathcal{S}$ and $\mathcal{S}'$ are independently generated from the following pre-change and post-change models, respectively,

$$\mathbf{Z}_i \sim p(\mathbf{z}_i|\theta),\ \mathbf{z}_i \in \mathcal{S}, \quad \text{and} \quad \mathbf{Z}'_j \sim p(\mathbf{z}'_i|\theta'),\ \mathbf{z}'_j \in \mathcal{S}'. \quad (1)$$

Our goal is to construct a computational efficient test to decide between the following two hypotheses:

$$\begin{aligned} H_0 &: (\theta, \theta') \in \chi_0 \triangleq \{(\theta, \theta')|\ \|\theta - \theta'\|_2 \le \rho\}, \\ H_1 &: (\theta, \theta') \in \chi_1 \triangleq \{(\theta, \theta')|\ \|\theta - \theta'\|_2 > \rho\}, \end{aligned} \quad (2)$$

where ρ is a constant determined by the specific applications.

Let $\delta : \mathcal{Z}^n \times \mathcal{Z}^{n'} \to \{0, 1\}$ denote the decision rule for the model change detection problem. Then the probabilities of false alarm and correct detection can be written as

$$\mathrm{P_F}(\delta, \theta, \theta') \triangleq \mathrm{P}_{(\theta,\theta')}\{\delta(S, S') = 1\}, \quad \forall(\theta, \theta') \in \chi_0, \quad (3)$$
$$\mathrm{P_D}(\delta, \theta, \theta') \triangleq \mathrm{P}_{(\theta,\theta')}\{\delta(S, S') = 1\}, \quad \forall(\theta, \theta') \in \chi_1, \quad (4)$$

where $\mathrm{P}_{(\theta,\theta')}$ denotes the probability measure for the data conditioned on the model parameter (θ, θ').

Note that in (2), both the null hypothesis and the alternative hypothesis are composite. We study the detection problem in the Neyman-Pearson setting:

$$\begin{aligned} \max_{\delta}\ & \mathrm{P_D}(\delta, \theta, \theta'), && \forall(\theta, \theta') \in \chi_1 \\ \text{s.t.}\ & \mathrm{P_F}(\delta, \theta, \theta') \le \alpha, && \forall(\theta, \theta') \in \chi_0. \end{aligned} \quad (5)$$

As seen in (5), our goal is to construct a test that maximizes the detection probability for all $(\theta, \theta') \in \chi_1$, and satisfies the false alarm constraint for all $(\theta, \theta') \in \chi_0$. The solution to (5) if it exists is said to be a uniformly most powerful (UMP) test.

Since $\mathbf{z}_i$ and $\mathbf{z}'_i$ are drawn i.i.d. from $p(\mathbf{z}_i|\theta)$ and $p(\mathbf{z}'_i|\theta')$, respectively, we can use

$$L(\theta) \triangleq -\sum_{i=1}^{n} \log p(\mathbf{z}_i|\theta), \quad L'(\theta) \triangleq -\sum_{i=1}^{n'} \log p(\mathbf{z}'_i|\theta) \quad (6)$$

to denote the negative log-likelihood functions with the pre-change dataset $\mathcal{S}$ and post-change dataset $\mathcal{S}'$, respectively. Then, the maximum likelihood estimates (MLE) of θ and θ' can be written as

$$\hat{\theta}_{\mathrm{ML}} \triangleq \operatorname{argmin} L(\theta), \qquad \hat{\theta}'_{\mathrm{ML}} \triangleq \operatorname{argmin} L'(\theta). \quad (7)$$

In addition, we denote the Hessian matrices of $L(\theta)$ and $L'(\theta)$ as $H(\theta) \triangleq \nabla^2_\theta L(\theta)$, and $H'(\theta) \triangleq \nabla^2_\theta L'(\theta)$.

3. EMPIRICAL DIFFERENCE TEST

3.1. Generalized Likelihood Ratio Test

In general, a UMP solution to the composite hypothesis testing problem in (5) may not exist, and may be difficult to find even if it exists. An alternative approach is to apply the GLRT. The generalized log-likelihood ratio (GLR) is given by

$$L_{\mathrm{G}}(\mathcal{S}, \mathcal{S}') \triangleq \log \frac{\max_{(\theta,\theta') \in \chi_1} \prod_{i=1}^{n} p(\mathbf{z}_i|\theta) \prod_{i=1}^{n'} p(\mathbf{z}'_i|\theta')}{\max_{(\theta,\theta') \in \chi_0} \prod_{i=1}^{n} p(\mathbf{z}_i|\theta) \prod_{i=1}^{n'} p(\mathbf{z}'_i|\theta')}. \quad (8)$$

If $L_{\mathrm{G}}(\mathcal{S}, \mathcal{S}')$ does not have point masses under either H_0 or H_1, the GLRT has the following structure

$$\delta_{\mathrm{GL}}(\mathcal{S}, \mathcal{S}') = \begin{cases} 1, & \text{if } L_{\mathrm{G}}(\mathcal{S}, \mathcal{S}') \ge \tau \\ 0, & \text{if } L_{\mathrm{G}}(\mathcal{S}, \mathcal{S}') < \tau, \end{cases} \quad (9)$$

where τ is the threshold for the GLR statistics determined by the false alarm constraint α.

For the conciseness, we define

$$\begin{aligned} (\hat{\theta}_1, \hat{\theta}'_1) &\triangleq \operatorname*{argmin}_{(\theta,\theta') \in \chi_1} L(\theta) + L'(\theta'), \\ (\hat{\theta}_0, \hat{\theta}'_0) &\triangleq \operatorname*{argmin}_{(\theta,\theta') \in \chi_0} L(\theta) + L'(\theta'). \end{aligned} \quad (10)$$

Then, the generalized log-likelihood ratio can be written as

$$L_{\mathrm{G}}(\mathcal{S}, \mathcal{S}') = L(\hat{\theta}_0) + L'(\hat{\theta}'_0) - L(\hat{\theta}_1) - L'(\hat{\theta}'_1). \quad (11)$$

The main difficulty in applying GLRT is that the minimizers $(\hat{\theta}_1, \hat{\theta}'_1)$ and $(\hat{\theta}_0, \hat{\theta}'_0)$ in (10) are hard to compute. In the following subsection, we propose an empirical difference test which approximates the GLRT and has reduced the computational complexity.

3.2. Empirical Difference Test

We need the following conditions to proceed with our analysis and establish the asymptotical normality of the MLEs [9].

Assumption 1 ***Regularity conditions for MLE***

1. ***Smoothness****: $L(\theta)$ and $L'(\theta)$ have first, second and third derivatives for all θ.*

2. ***Strong Convexity****: For all θ, $H(\theta)$ and $H'(\theta)$ are positive definite.*

3. ***Boundedness****: For all θ, the largest eigenvalues of $H(\theta)$ and $H'(\theta)$ are upper bounded by λ_M.*

We note that the MLEs $(\hat{\theta}_{\mathrm{ML}}, \hat{\theta}'_{\mathrm{ML}})$ belong to either χ_0 or χ_1. If $(\hat{\theta}_{\mathrm{ML}}, \hat{\theta}'_{\mathrm{ML}}) \in \chi_1$, i.e., $(\hat{\theta}_1, \hat{\theta}'_1) = (\hat{\theta}_{\mathrm{ML}}, \hat{\theta}'_{\mathrm{ML}})$, we have $L_{\mathrm{G}}(\mathcal{S}, \mathcal{S}') = L(\hat{\theta}_0) - L(\hat{\theta}_{\mathrm{ML}}) + L'(\hat{\theta}'_0) - L'(\hat{\theta}'_{\mathrm{ML}}) > 0$.

In addition, the worst-case false alarm probability of GLRT is given by $\max_{(\theta,\theta' \in \chi_0)} \mathrm{P}_{(\theta,\theta')}\{L_{\mathrm{G}}(\mathcal{S}, \mathcal{S}') \ge \tau\}$, which we wish to upper bounded by α. Note that $L_{\mathrm{G}}(\mathcal{S}, \mathcal{S}') > 0$ when $(\hat{\theta}_{\mathrm{ML}}, \hat{\theta}'_{\mathrm{ML}}) \in \chi_1$ holds. In the following, we focus on the case where $\alpha < \max_{(\theta,\theta' \in \chi_0)} \mathrm{P}_{(\theta,\theta')}\{L_{\mathrm{G}}(\mathcal{S}, \mathcal{S}') \ge 0\}$, i.e., a relatively small false alarm constraint α. Thus, we just need to study the false alarm probability of GLRT when $(\hat{\theta}_{\mathrm{ML}}, \hat{\theta}'_{\mathrm{ML}}) \in \chi_1$ and $\tau > 0$.

Given $(\hat{\theta}_{\mathrm{ML}}, \hat{\theta}'_{\mathrm{ML}}) \in \chi_1$, it is difficult to solve for $(\hat{\theta}_0, \hat{\theta}'_0)$ in (10) exactly. However, we can construct an upper bound for the GLR by approximating $(\hat{\theta}_0, \hat{\theta}'_0)$ using a linear combination of $(\hat{\theta}_{\mathrm{ML}}, \hat{\theta}'_{\mathrm{ML}})$. Let $\Delta\hat{\theta} \triangleq \hat{\theta}'_{\mathrm{ML}} - \hat{\theta}_{\mathrm{ML}}$. Then $\|\Delta\hat{\theta}\|_2 > \rho$,

$$\widetilde{\theta}_0 = \hat{\theta}_{\mathrm{ML}} + \frac{\mu\Delta\hat{\theta}}{\|\Delta\hat{\theta}\|_2}, \quad \widetilde{\theta}'_0 = \hat{\theta}_{\mathrm{ML}} + \frac{(\mu+\rho)\Delta\hat{\theta}}{\|\Delta\hat{\theta}\|_2}, \tag{12}$$

where $\mu \in [0, \|\Delta\hat{\theta}\|_2 - \rho]$ denotes the distance between $\widetilde{\theta}_0$ and $\hat{\theta}_{\mathrm{ML}}$. It can be verified that $(\widetilde{\theta}_0, \widetilde{\theta}'_0) \in \chi_0$. Then, the GLR in (11) can be upper bounded as

$$\begin{aligned}
&L_{\mathrm{G}}(\mathcal{S}, \mathcal{S}') = L(\hat{\theta}_0) + L'(\hat{\theta}'_0) - L(\hat{\theta}_1) - L(\hat{\theta}'_1) \\
&\le L(\widetilde{\theta}_0) + L'(\widetilde{\theta}'_0) - L(\hat{\theta}_1) - L(\hat{\theta}'_1) \\
&\overset{(a)}{=} \frac{1}{2}(\hat{\theta}_1 - \widetilde{\theta}_0)^\top H(\widetilde{\theta})(\hat{\theta}_1 - \widetilde{\theta}_0) + \frac{1}{2}(\hat{\theta}'_1 - \widetilde{\theta}'_0)^\top H'(\widetilde{\theta}')(\hat{\theta}'_1 - \widetilde{\theta}'_0) \\
&= \frac{\mu^2}{2}\frac{\Delta\hat{\theta}^\top}{\|\Delta\hat{\theta}\|_2} H(\widetilde{\theta}) \frac{\Delta\hat{\theta}}{\|\Delta\hat{\theta}\|_2} \\
&\quad + \frac{(\|\Delta\hat{\theta}\|_2 - (\mu+\rho))^2}{2}\frac{\Delta\hat{\theta}^\top}{\|\Delta\hat{\theta}\|_2} H'(\widetilde{\theta}') \frac{\Delta\hat{\theta}}{\|\Delta\hat{\theta}\|_2} \qquad (13) \\
&\overset{(b)}{\le} \frac{\mu^2}{2\sigma^2}\lambda_{\max}(H(\widetilde{\theta})) + \frac{(\|\Delta\hat{\theta}\|_2 - (\mu+\rho))^2}{2\sigma^2}\lambda_{\max}(H'(\widetilde{\theta}')),
\end{aligned}$$

where (a) follows from the Taylor's Theorem, $\widetilde{\theta}$ and $\widetilde{\theta}'$ denote the parameters in the corresponding remainders; and (b) follows from the fact that $H(\widetilde{\theta})$ and $H'(\widetilde{\theta}')$ are positive definite and $\frac{\Delta\hat{\theta}^\top}{\|\Delta\hat{\theta}\|_2}$ is a unit vector. Note that $\lambda_{\max}(H(\widetilde{\theta}))$ and $\lambda_{\max}(H'(\widetilde{\theta}'))$ are bounded by λ_M in Assumption 1. Hence,

$$\begin{aligned}
&\mathrm{P}_{\mathrm{F}}(\delta_{\mathrm{GL}}) = \mathrm{P}_{(\theta,\theta')}\{L_{\mathrm{G}}(\mathcal{S}, \mathcal{S}') \ge \tau\} \\
&\le \mathrm{P}_{(\theta,\theta')}\Big\{\frac{\mu^2}{2\sigma^2}\lambda_M + \frac{(\|\Delta\hat{\theta}\|_2 - (\mu+\rho))^2}{2\sigma^2}\lambda_M \ge \tau\Big\} \\
&= \mathrm{P}_{(\theta,\theta')}\{\|\Delta\hat{\theta}\|_2 \ge \eta\}, \qquad (14)
\end{aligned}$$

for $(\theta, \theta') \in \chi_0$. The false alarm probability of GLRT can be upper bounded by the probability that the *empirical difference* $\|\Delta\hat{\theta}\|_2$ is larger than another threshold η. Note that the threshold η can be set by letting $\mathrm{P}_{(\theta,\theta')}\{\|\Delta\hat{\theta}\|_2 \ge \eta\} \le \alpha$ for all $(\theta, \theta') \in \chi_0$, which is independent of the unknown quantities μ and λ_M.

Thus, we propose the following empirical difference test with the following structure to approximate the GLRT,

$$\delta_{\mathrm{ED}} = \begin{cases} 1, & \text{if } \|\Delta\hat{\theta}\|_2 \ge \eta \\ 0, & \text{if } \|\Delta\hat{\theta}\|_2 < \eta. \end{cases} \tag{15}$$

The benefits for using δ_{ED} are two-fold: 1) Instead of constructing the more complicated GLR statistics, our EDT only requires the computation of the empirical difference $\Delta\hat{\theta}$ between the MLEs, which is more tractable in practice. 2) The distribution of the empirical difference $\Delta\hat{\theta}$ is asymptotically Gaussian, which facilitates the setting of the threshold η to meet the false alarm constraint α.

4. APPROXIMATION FOR SETTING TEST THRESHOLD

In this subsection, we provide a method based on a χ^2 approximation [10] to set the threshold η in the EDT.

Since $\hat{\theta}_{\mathrm{ML}}$ and $\hat{\theta}'_{\mathrm{ML}}$ are the MLEs of θ and θ' with n and n' samples, respectively, we have

$$\sqrt{n}(\hat{\theta}_{\mathrm{ML}} - \theta) \xrightarrow{d.} \mathcal{N}(0, I_\theta^{-1}), \ \sqrt{n'}(\hat{\theta}'_{\mathrm{ML}} - \theta') \xrightarrow{d.} \mathcal{N}(0, I_{\theta'}^{-1}),$$

from the asymptotical normality of MLE [9], where I_θ denotes the Fisher information matrix of the probabilistic model $p(\mathbf{z}|\theta)$. Thus, we can approximate the distribution of $\Delta\hat{\theta}$ using a Gaussian distribution $\mathcal{N}(\theta' - \theta, \Sigma_{\Delta\theta})$, where $\Sigma_{\Delta\theta} \triangleq \frac{I_\theta^{-1}}{n} + \frac{I_{\theta'}^{-1}}{n'}$. In practice, I_θ and $I_{\theta'}$ can be estimated by replacing θ and θ' with the corresponding MLEs $\hat{\theta}_{\mathrm{ML}}$ and $\hat{\theta}'_{\mathrm{ML}}$, respectively.

To satisfy the false alarm constraint in (5), we need to set the threshold η_α based on the following equation in the EDT,

$$\max_{\theta,\theta' \in \chi_0} \mathrm{P}_{(\theta,\theta')}\{\|\Delta\hat{\theta}\|^2 \ge \eta_\alpha^2\} = \alpha. \tag{16}$$

The following theorem characterizes the distribution of $\|\Delta\hat{\theta}\|^2$ that results from the Gaussian approximation.

Theorem 1 *Suppose* $\Delta\theta \sim \mathcal{N}(\theta' - \theta, \Sigma_{\Delta\theta})$*, and the covariance matrix* $\Sigma_{\Delta\theta}$ *has the eigen-decomposition* $\Sigma_{\Delta\theta} = P^\top \Lambda P$*, where* $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$ *contains all the eigenvalues, and* P *is an orthogonal matrix. Then,*

$$\|\Delta\theta\|^2 \overset{d.}{=} \sum_{i=1}^{d} \lambda_i (U_i + \mathbf{b}_i)^2, \tag{17}$$

where $U_i \sim \mathcal{N}(0, 1)$*, and* $\mathbf{b} = (\sqrt{\Lambda})^{-1}(\theta' - \theta)$.

The distribution of $\|\Delta\theta\|^2$ is a linear combination of independent non-central chi-squared random variables with degree of freedom of one, which does not have a simple closed form [11]. We therefore propose the following approximation method to set the threshold in the EDT. Note that

$$\begin{aligned}
&\mathrm{P}_{(\theta,\theta')}\Big\{\sum_{i=1}^{d} \lambda_i (U_i + \mathbf{b}_i)^2 \ge \eta^2\Big\} \\
&\quad \le \mathrm{P}_{(\theta,\theta')}\Big\{\sum_{i=1}^{d} (U_i + \mathbf{b}_i)^2 \ge \eta^2 / \lambda_{\max}(\Sigma_{\Delta\theta})\Big\}, \qquad (18)
\end{aligned}$$

for $(\theta, \theta') \in \chi_0$, and $\sum_{i=1}^{d}(U_i + \mathbf{b}_i)^2$ is a non-central chi-squared $\chi^2(k, \gamma)$ random variable with degrees of freedom $k = d$, and non-centrality parameter $\gamma = \sum_{i=1}^{d} \mathbf{b}_i^2 \le \rho^2 / \lambda_{\min}(\Sigma_{\Delta\theta})$, where the inequality follows from the fact $\|\theta' - \theta\|_2 \le \rho$ under H_0. Thus,

$$\begin{aligned}
&\max_{\theta,\theta' \in \chi_0} \mathrm{P}_{(\theta,\theta')}\{\|\Delta\theta\|_2^2 \ge \eta^2\} \\
&\quad \le \max_{\theta,\theta' \in \chi_0} \mathrm{P}\Big\{\chi^2(d, \sum_{i=1}^{d} b_i^2) \ge \eta^2 / \lambda_{\max}(\Sigma_{\Delta\theta})\Big\}. \qquad (19)
\end{aligned}$$

We can set the threshold $\widetilde{\eta}_\alpha$ with the χ^2 approximation [10] using the following equation,

$$\mathrm{P}\{\chi^2(d, \rho^2/\lambda_{\min}(\Sigma_{\Delta\theta})) \geq \widetilde{\eta}_\alpha^2/\lambda_{\max}(\Sigma_{\Delta\theta})\} = \alpha \qquad (20)$$

to ensure that the false alarm probability is bounded by α.

5. NUMERICAL RESULTS

In this section, we evaluate the performance of the proposed empirical difference test δ_{ED} in linear regression and logistic regression models.

Linear regression model: The datasets $\mathcal{S}$ and $\mathcal{S}'$ are generated from the linear model $\mathbf{y} = X\theta + \xi$, where $X \in \mathbb{R}^{n\times d}$ denotes the input variable, $\mathbf{y} \in \mathbb{R}^n$ denotes the response variable and $\theta \in \mathbb{R}^d$ denotes the weight vector. We assume that all the elements in noises $\xi \in \mathbb{R}^n$ are i.i.d. zero mean Gaussian random variables generated from $\mathcal{N}(0, \sigma^2)$. Then, the Fisher information matrix $I_\theta = XX^\top/\sigma^2$ is independent of θ. In the simulations, we set the dimension $d = 10$, the number of samples $n = n' = 40$, $\sigma^2 = 1$ and $\rho = 1$.

Logistic regression model: The datasets $\mathcal{S}$ and $\mathcal{S}'$ are generated from the following logistic model

$$p(y_i|\mathbf{x}_i, \theta) = \frac{1}{1 + \exp(-y_i\mathbf{x}_i^\top\theta)}, \ \forall(\mathbf{x}_i, y_i) \in \mathcal{S}, \qquad (21)$$

where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the feature vector, $y_i \in \{\pm1\}$ denotes the label, and $\theta \in \mathbb{R}^d, \|\theta\|_2 = 1$ is the normalized model parameter vector. Then, the Fisher information matrix

$$I_\theta = \mathbb{E}_{\mathbf{x}}\Big[\frac{1}{1 + \exp(\mathbf{x}_i^\top\theta)}\frac{1}{1 + \exp(-\mathbf{x}_i^\top\theta)}\mathbf{x}_i\mathbf{x}_i^\top\Big]. \qquad (22)$$

In the simulations, we choose dimension $d = 5$, the number of samples $n = n' = 60$, and set ρ such that the angle between θ and θ' is $\frac{\pi}{4}$.

To illustrate the performance of the proposed algorithms, we plot the probability $\mathrm{P}_{(\theta,\theta')}\{\delta = 1\}$ as a function of $\|\theta - \theta'\|_2$ in all three figures, where the normalized model change $\|\theta - \theta'\|/\rho$ ranges from 0 to 2. Note that when $\|\theta - \theta'\|_2 < \rho$, i.e., $(\theta, \theta') \in \chi_0$, $\mathrm{P}_{(\theta,\theta')}\{\delta = 1\}$ denotes the false alarm probability $\mathrm{P}_{\mathrm{F}}(\delta)$ (in the left side of the figures). In contrast, when $\|\theta - \theta'\|_2 > \rho$, $(\theta, \theta') \in \chi_1$ and $\mathrm{P}_{(\theta,\theta')}\{\delta = 1\}$ denotes the detection probability $\mathrm{P}_{\mathrm{D}}(\delta)$ (in the right side of the figures). Thus, the plot of $\mathrm{P}_{(\theta,\theta')}\{\delta = 1\}$ provides us with an illustration of the test performance under both hypotheses with different model parameters.

To verify the approximation of the GLRT with the proposed EDT, we first compare the performance of these tests for the linear regression model (the GLRT is not computationally feasible for logistic regression) for two values of the false alarm constraint $\alpha = 0.1$ and $\alpha = 0.3$. The thresholds of these tests η_α are set using 1000 runs of Monte-Carlo simulations such that the false alarm probabilities are equal to α as in (16). It is shown in Fig. 1 that the difference between the performance of EDT and that of GLRT is negligible with only $n = n' = 40$ samples, which justifies the use of EDT. We note that when $\|\theta - \theta'\|/\rho = 1$, it is impossible to distinguish H_0 and H_1 even if the number of samples n and n' go to infinity, i.e., the probabilities of false alarm and detection are both equal to α in this case.

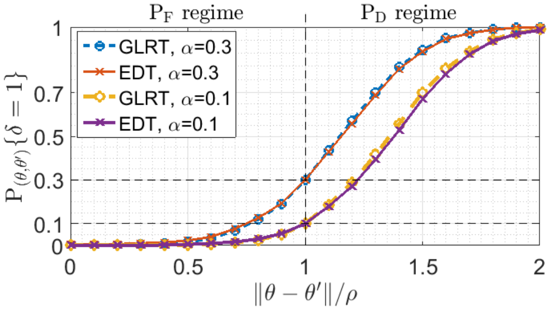

Fig. 1. Comparison of the performances of the GLRT and EDT for the linear regression model.

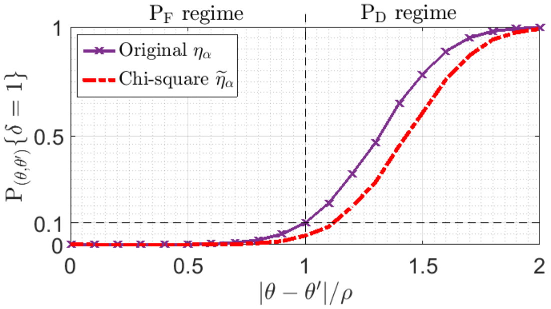

Fig. 2. Comparison of the performance of EDT with the threshold η_α and the χ^2 approximation $\widetilde{\eta}_\alpha$, for the linear regression model with $\alpha = 0.1$.

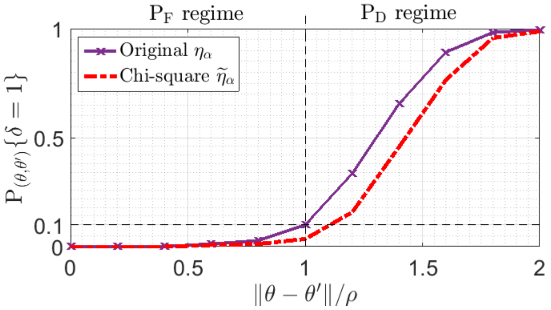

Fig. 3. Comparison of the performance of EDT with the threshold η_α and the χ^2 approximation $\widetilde{\eta}_\alpha$, for the logistic regression model with $\alpha = 0.1$.

Fig. 2 and Fig. 3 compare the performance of EDT with the threshold η_α computed by 1000 runs of Monte-Carlo simulations in (16), and the threshold $\widetilde{\eta}_\alpha$ set by the proposed χ^2 approximation in (20), respectively, when $\alpha = 0.1$. It can be observed that in both linear regression and logistic regression cases, the non-central chi-squared approximation in (20) provides conservative estimates of the test thresholds η, thereby ensuring that the false alarm constraint is met.

6. REFERENCES

[1] C. Wilson, V. V Veeravalli, and A. Nedich, “Adaptive sequential stochastic optimization,” *IEEE Transactions on Automatic Control*, 2018.

[2] C. Wilson and V. V Veeravalli, “Adaptive sequential optimization with applications to machine learning,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 2642–2646.

[3] Y. Bu, J. Lu, and V. V Veeravalli, “Active and adaptive sequential learning,” *arXiv preprint arXiv:1805.11710*, 2018.

[4] M. Elahi, F. Ricci, and N. Rubens, “A survey of active learning in collaborative filtering recommender systems,” *Computer Science Review*, vol. 20, pp. 29–50, 2016.

[5] N. Rubens, M. Elahi, M. Sugiyama, and D. Kaplan, “Active learning in recommender systems,” in *Recommender Systems Handbook*, pp. 809–846. Springer, 2015.

[6] S. J Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[7] J. Geng, B. Zhang, L. M Huie, and L. Lai, “Online change detection of linear regression models,” in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4910–4914.

[8] S. Zou, G. Fellouris, and V. V Veeravalli, “Quickest change detection under transient dynamics: Theory and asymptotic analysis,” *arXiv preprint arXiv:1711.02186*, 2017.

[9] A. W. Van der Vaart, *Asymptotic statistics*, Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 2000.

[10] A. DasGupta, *Asymptotic Theory of Statistics and Probability*, Springer, 2008.

[11] S. J. Press, “Linear combinations of non-central chisquare variates,” *The Annals of Mathematical Statistics*, pp. 480–487, 1966.