Model Change Detection with Application to Machine Learning

Yuheng Bu†, Jiaxun Lu‡, and Venugopal V. Veeravalli†

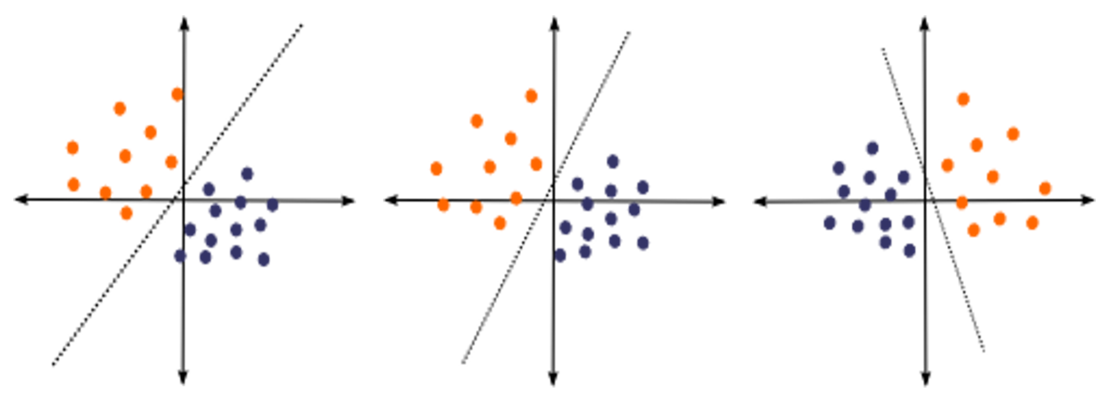
† University of Illinois at Urbana-Champaign ‡ Tsinghua University

ILLINOIS Electrical & Computer Engineering COLLEGE OF ENGINEERING

1. Introduction

In adaptive sequential learning

- Models learned in previous steps are used adaptively to improve accuracy in next steps
- Adapting to previous model that is significantly different from the current one could deteriorate performance
- Detect significant model change with samples



$t=1$ $t=2$ $t=3$

2. Problem Model

- Two datasets $\mathcal{S} = \{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$ and $\mathcal{S}' = \{\mathbf{z}'_1, \cdots, \mathbf{z}'_{n'}\}$ from some instance space $\mathcal{Z}$
- Parameterized family of distribution models $\mathcal{M} = \{p(\mathbf{z}|\theta), \theta \in \mathbb{R}^d\}$
- Unknown parameters $\theta, \theta' \in \mathbb{R}^d$, such that

Pre-change model **Post-change model**

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

Goal: construct efficient test $\delta : \mathcal{Z}^n \times \mathcal{Z}^{n'} \to \{0,1\}$ to decide between following hypotheses:

$$H_0 : (\theta, \theta') \in \chi_0 \triangleq \{(\theta, \theta') | \, \|\theta - \theta'\|_2 \le \rho\},$$
$$H_1 : (\theta, \theta') \in \chi_1 \triangleq \{(\theta, \theta') | \, \|\theta - \theta'\|_2 > \rho\},$$

where ρ is a constant determined by application.

Reference

[1] **Y. Bu**, J. Lu, and V.V. Veeravalli, Model change detection with application to machine learning, ICASSP 2019

[2] C. Wilson and V. V. Veeravalli and A.Nedich, Adaptive Sequential Stochastic Optimization, IEEE Transaction on Automatic Control, 2018

3. Notations

Probabilities of false alarm and detection

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

Neyman-Pearson setting:

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

The solution is said to be a uniformly most powerful (UMP) test. Denote

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

Maximum likelihood estimates (MLE) of θ and θ'

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

4. Empirical Difference Test (EDT)

In general, UMP test may not exist. One alternate is to use generalized likelihood ratio test (GLRT).

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

- Main difficulty of GLRT is that optimization problem is computationally hard to solve.
- We show that false alarm probability of GLRT can be upper bounded by the probability that norm of *empirical difference*

$$\Delta\hat{\theta} = \hat{\theta}_{\mathrm{ML}} - \hat{\theta}'_{\mathrm{ML}}$$

  is larger than another threshold.
- We propose *empirical difference test* (EDT) to approximate GLRT

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

Threshold η_α is set by

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

5. Approximation for setting test threshold

How to set threshold for EDT?

- $\hat{\theta}_{\mathrm{ML}}$ and $\hat{\theta}'_{\mathrm{ML}}$ are MLEs of θ and θ'
- Under regularity conditions, asymptotical normality of MLE gives

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

- Approximating the distribution of $\Delta\hat{\theta}$ with

$$\mathcal{N}(\theta' - \theta, \Sigma_{\Delta\theta}), \quad \Sigma_{\Delta\theta} \triangleq \frac{I_\theta^{-1}}{n} + \frac{I_{\theta'}^{-1}}{n'}$$

  I_θ denotes Fisher information matrix
- In practice, I_θ and $I_{\theta'}$ can be estimated by replacing θ and θ' with corresponding MLEs

Theorem: Suppose $\Delta\hat{\theta} \sim \mathcal{N}(\theta' - \theta, \Sigma_{\Delta\theta})$, and $\Sigma_{\Delta\theta}$ has eigen-decomposition $\Sigma_{\Delta\theta} = P^\top \Lambda P$, where $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$ contains all eigen-values, and P is orthogonal. Then,

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

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

Main difficulties:

- Distribution of $\|\Delta\hat{\theta}\|^2$ is linear combination of independent non-central chi-squared random variables with degree of freedom of one
- **No** simple closed form

Using χ^2 approximation, we show that false alarm probability of EDT can be upper bounded by

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

Set threshold $\tilde{\eta}_\alpha$ with χ^2 approximation

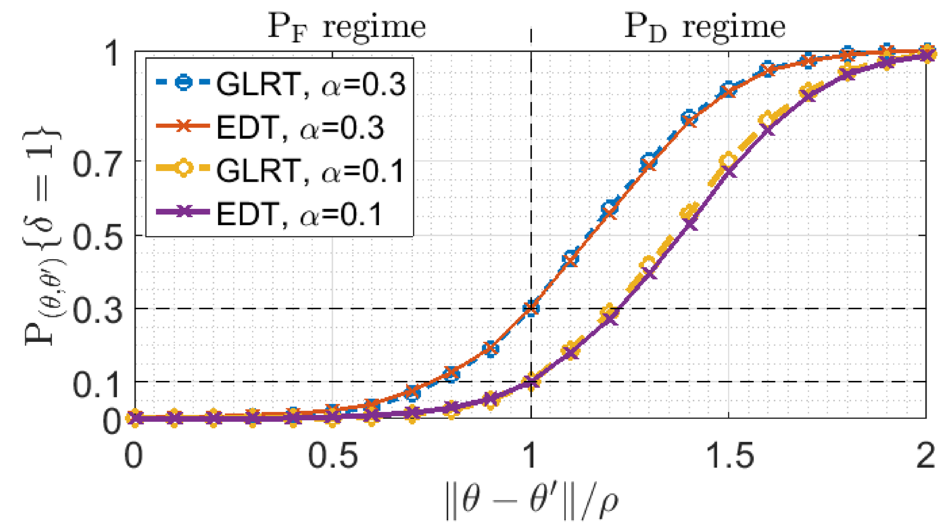
$$\mathrm{P}\{\chi^2(d, \rho^2/\lambda_{\min}(\Sigma_{\Delta\theta})) \ge \tilde{\eta}_\alpha^2/\lambda_{\max}(\Sigma_{\Delta\theta})\} = \alpha$$

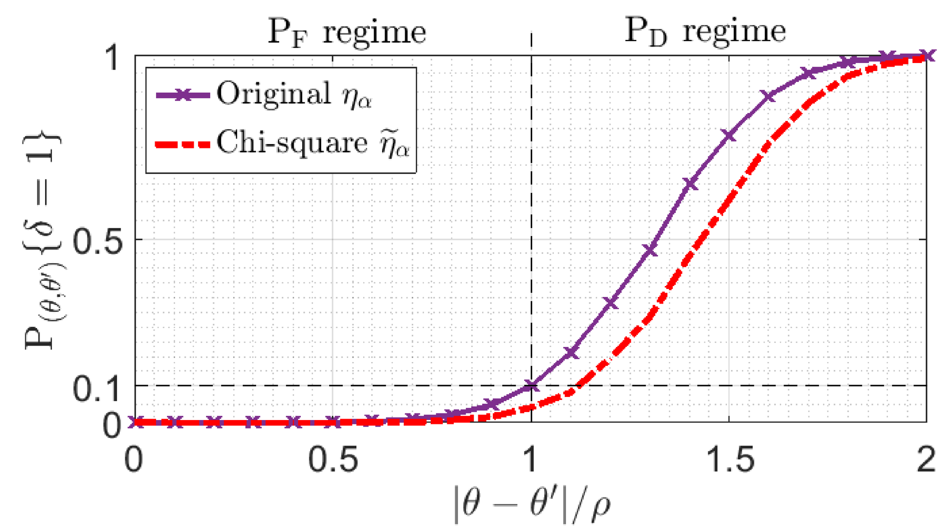
to ensure false alarm probability is bounded by α.

6. Numerical Results

Linear regression model:

- Linear regression $\mathbf{y} = X\theta + \xi$
- $X \in \mathbb{R}^{n\times d}$, $\mathbf{y} \in \mathbb{R}^n$, $\theta \in \mathbb{R}^d$
- Noise: $\xi \in \mathcal{N}(0, I_n)$
- Dimension $d = 10$, $n = n' = 40$, $\rho = 1$





Logistic regression model:

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

- $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{\pm 1\}$
- Normalized parameter $\theta \in \mathbb{R}^d$, $\|\theta\|_2 = 1$
- Dimension $d = 5$, $n = n' = 60$
- Set ρ such that angle between θ and θ' is $\frac{\pi}{4}$