Quickest Detection in High-Dimensional Linear Regression Models via Implicit Regularization

Qunzhi Xu*, Yi Yu† and Yajun Mei*‡
* : School of Industrial and Systems Engineering, Georgia Institute of Technology, {xuqunzhi, ymei3}@gatech.edu
† : Department of Statistics, University of Warwick, yi.yu.2@warwick.ac.uk
‡ : Department of Biostatistics, School of Global Public Health, New York University

***Abstract*—In this paper, we consider the quickest detection problem in high-dimensional streaming data, where the unknown regression coefficients might change at some unknown time. We propose a quickest detection algorithm based on the implicit regularization algorithm via gradient descent, and provide theoretical guarantees on the average run length to false alarm and detection delay. Numerical studies are conducted to validate the theoretical results.**

I. Introduction

In this paper, we tackle the problem of detecting changes in high-dimensional linear regression models - one of the most fundamental predictive models. Our approach is inspired by recent research on (deep) neural networks, where optimization algorithms such as (stochastic) gradient descent hold *implicit regularization* properties, see [1], [2], etc. By treating linear regression models as two-layers neural networks, we develop efficient implicit regularization-based quickest detection for linear regression models. To the best of our knowledge, this is the first to apply implicit regularization in the text of quickest detection, and our ideas can be easily extended to the context of monitoring other advanced predictive machine learning or artificial intelligent algorithms.

To be specific, we assume that we observe a sequence $\{(y_t, X_t)\}_{t\in\mathbb{Z}_+} \subset \mathbb{R}^m \times \mathbb{R}^{m\times p}$ over time $t \in \mathbb{Z}_+$ which can be modeled as

$$y_t = X_t\beta_t + \epsilon_t, \tag{1}$$

where $\beta_t \in \mathbb{R}^p$ is a p-dimensional vector. The system is initially in control and the parameter vector takes a constant value $\beta_t = \beta_0$. At some unknown time $\nu \in \mathbb{Z}_+$, an event occurs and changes the regression coefficient vector to $\beta_1 \in \mathbb{R}^p$, $\beta_1 \neq \beta_0$. In the literature, such time ν is often called the "change point". The primary goal is to develop an efficient algorithm to raise an alarm as quickly as possible once it occurs at change point ν based on the observed data sequence $\{(y_t, X_t)\}_{t\in\mathbb{Z}_+}$.

Research on online/sequential monitoring of change points in high-dimensional linear model has been studied in the statistical and engineering literature, e.g., [3], [4], but most through *explicit regularization* methods such as LASSO proposed in [5]. For instance, [6] developed a LASSO-based multivariate statistical process control (SPC) methodology. [7] developed residual-based detection statistics via LASSO estimator. Here we take a different approach by adopting the implicit regularization method, which might allow us to handle more complicated models or algorithms.

The rest of the paper is organized as follows: in Section II we formulate the problem and review the implicit regularization methods in linear regression model. In Section III we propose the general algorithm and the efficient implementation via implicit regularization. In Section IV we develop the theoretical properties of the proposed algorithm and in Section V we conduct several numerical studies to validate our results. Finally in Section VI we provide some technical details.

II. Problem Formulation and Background

A. Mathematical Formulation

Assume that we observe a sequence of data streams $\{(y_t, X_t)\}_{t\in\mathbb{Z}_+} \subset \mathbb{R}^m \times \mathbb{R}^{m\times p}$ from model (1), where the noise vector $\epsilon_t \in \mathbb{R}^m \sim N(0, \sigma^2 I)$, and the regression coefficient β_t change from a pre-specified vector β_0 to an unknown vector $\beta^* \neq \beta_0$ at some unknown change point ν, i.e.

$$\beta_t = \begin{cases} \beta_0, & t \leq \nu, \\ \beta^*, & t > \nu. \end{cases} \tag{2}$$

The known value of β_0 is reasonable in many applications such as the quality control in manufacturing engineering. Without loss of generality, we assume $\beta_0 = 0$ since we can monitor $y_t - X_t\beta_0$ instead of y_t.

We consider the random design case, where each component $(X_t)_{i,j}$ of the matrix X_t is assumed to be independent and identically distributed random variables with the standard Gaussian distribution (i.e., $(X_t)_{i,j} \sim N(0,1)$), for all $i = 1, \cdots, m, j = 1, \cdots, p$ and $t = 1, \cdots,$. In addition, we assume that the change is entrywise sparse, and define the sparsity $s^* = \|\beta^* - \beta_0\|_0$, where $\|\cdot\|_0$ denotes the L_0 norm.

Our goal is to develop an efficient algorithm to detect the change based on the observed data $\{(y_t, X_t)\}_{t\in\mathbb{Z}_+}$ as quickly as possible. An algorithm for quickest detection problem can be characterized by a stopping time T with respect to the observed data sequence, where $T = n$ means that we raise a global alarm at time n. Denote by P_ν and E_ν the probability measure and expectation when change occurs at time t. Denote by P_∞ and E_∞ the probability measure and expectation when there are no changes, or equivalently, when the change occurs at time ∞. Motivated by [8], the detection delay of a stopping

time T can be evaluated by the following worst case detection delay conditioned on $T > \nu$:

$$D(T) = \sup_{\nu \geq 1} \mathsf{E}_\nu[T - \nu \mid T > \nu], \tag{3}$$

subject to the average run length (ARL) to false alarm control:

$$\mathsf{E}_\infty[T] \geq \gamma, \tag{4}$$

for some pre-specified $\gamma > 1$.

Had we known the true value of the post-change vector β^*, the problem can be solved via classical CUSUM procedure in Page [9]. Mathematically, denote by $f_{X\beta^*}$ as the probability density function (pdf) of $N(X\beta^*, \sigma^2 I)$, we define the following detection statistics:

$$W_t = \max\left\{W_{t-1} + \log(\frac{f_{X_t\beta^*}(y_t)}{f_0(y_t)}), 0\right\}, \tag{5}$$

for $t \geq 1$ with $W_0 = 0$. The corresponding CUSUM stopping time is then defined as

$$T_{\text{CUSUM}} = \inf\{t > 0 : W_t \geq A\}, \tag{6}$$

for some pre-specified constant $A > 0$.

When the post-change regression coefficient β^* is unknown, an intuitive idea would be to construct an estimator $\hat{\beta}_t$ based on the historical data and plug it into the standard CUSUM statistics in (5) for detection. However, there are two main challenges: (1) the subset of data to be used for the estimation is unclear. Due to the unknown change point ν, the data used might be a mixture model from pre-change and post-change scenarios. This results in a potentially large bias of estimators and detection delay. (2) Efficient and accurate estimators remain unclear. We therefore review two signal recovery methods in high-dimensional linear regression model in the next subsection.

B. Review of Signal Recovery Methods in Linear Models

Consider data $(y, X) = (y_t, X_t) \subset \mathbb{R}^m \times \mathbb{R}^{m\times p}$ generated from model (1). A standard approach when p does not diverge with m is to estimate β^* via least squares estimators (LSE), which aims to find β that minimizes the residual sum squares (RSS). The optimization problem can be efficiently solved via the gradient descent algorithm.

In the high-dimensional scenario when p is diverging with m, many efficient algorithms have been developed, such as Lasso proposed in [5]. Researchers recently find out that combining the over-parametrization with gradient descent can lead to a sparse solution that achieves the minimax rate, which is also known as the implicit regularization algorithm in linear models, see [10], and [11].

To be more specific, for any vector $\beta \in \mathbb{R}^p$, we over-parameterize β by two vectors $u, v \in \mathbb{R}^p$:

$$\beta = u \circ u - v \circ v,$$

where $\circ$ is the Hadamard product that denotes the pointwise multiplication. Denote by $\|\cdot\|_2$ the L_2 norm, the RSS becomes

$$L(u, v) = \frac{1}{m} \|X(u \circ u - v \circ v) - y\|_2^2, \tag{7}$$

under the over-parameterization. We then apply the gradient descent algorithm to (7) to recursively update u and v by

$$\begin{aligned}
\beta_\ell &= u_\ell \circ u_\ell - v_\ell \circ v_\ell, \\
u_{\ell+1} &= u_\ell - 4\eta\, u_\ell \circ \left[\frac{1}{n} X^\top \{X\beta_\ell - y\}\right], \\
v_{\ell+1} &= v_\ell + 4\eta\, v_\ell \circ \left[\frac{1}{n} X^\top \{X\beta_\ell - y\}\right].
\end{aligned}$$

for $\ell = 1, \cdots, L_{\max}$, where $L_{\max}$ is the pre-specified maximum iteration number. The estimator can then be written as $\hat{\beta} = u_{L_{\max}} \circ u_{L_{\max}} - v_{L_{\max}} \circ v_{L_{\max}}$. Here the gradient descent is initialized by $u_{0,i}, v_{0,i} \sim \text{Unif}[-\alpha, \alpha]$ for some $\alpha > 0$ and $i = 1, \cdots, p$. Parameters including step size η, iteration number $L_{\max}$, and magnitude α need to be properly selected to ensure the optimal rate of implicit regularization algorithm in high-dimensional linear models.

III. Implicit Regularization-Based Quickest Detection

Our proposed stopping time T_{IR} contains three key components: the estimators $\hat{\beta}_t$, the monitoring statistics W_t, and the candidate change point $M(t)$. At a high level, for each time instant t, we construct the estimators $\hat{\beta}_t$ based on the data between the candidate change point $M(t)$ and the current time t. Then we update the statistics W_t based on $\hat{\beta}_t$ and choose to re-set or keep the candidate change point $M(t)$ based on the value of W_t. For the better presentation, we define the three components separately in three subsections.

A. Estimators $\hat{\beta}_t$

Let us begin with the construction of the estimator $\hat{\beta}_t$ of the true coefficient β^*. At each time t, if the candidate change point $M(t) = t - 1$, we directly set $\hat{\beta}_t = 0 \in \mathbb{R}^p$, which is exactly the value of pre-change coefficient. Otherwise if $M(t) < t - 1$, we consider the time window between the candidate change point $M(t)$ and current time t: $[M(t) + 1, t-1]$. Denote by $y_{M(t)+1,t-1}, X_{M(t)+1,t-1}$ the aggregation of the observed data in this time window:

$$\begin{aligned}
y_{M(t)+1,t-1} &= [y_{M(t)+1}^\top, \cdots, y_{t-1}^\top]^\top, \\
X_{M(t)+1,t-1} &= [X_{M(t)+1}^\top, \cdots, X_{t-1}^\top]^\top.
\end{aligned} \tag{8}$$

We define the proposed estimator $\hat{\beta}_t$ in three steps:

1) We introduce six tuning parameters: $\alpha_0, \eta_0, c_L > 0$ for the implementation of implicit regularization algorithm, and $s, c, C > 0$ for the truncation of the estimator.
2) we implement the implicit regularization algorithm in (8) to the aggregated data $(y_{M(t)+1,t-1}, X_{M(t)+1,t-1})$ to obtain an initial estimator $\tilde{\beta}_t$ with the initial value α_t, step size η_t and the maximum iteration number L_t defined in:

$$\begin{aligned}
\alpha_t &= \frac{\alpha_0}{\sqrt{t - M(t) - 1}}, \\
\eta_t &= \eta_0, \\
L_t &= \frac{c_L (t - M(t) - 1)^{1/4}}{\eta_t \sigma \sqrt{\log p/m}} \log \frac{1}{\alpha_t}.
\end{aligned}$$

3) We adjust the estimator $\tilde{\beta}_t$ to obtain the final estimator $\hat{\beta}_t$ by truncating the s largest components of $\tilde{\beta}_t$ with lower threshold c and upper threshold C, and setting the remaining $p-s$ components to 0. Mathematically, if we denote by $\hat{S}_t \subset \{1, \cdots, p\}$ the index of s largest components at time t, then

$$(\hat{\beta}_t)_i = 0, \text{ if } i \notin \hat{S}_t,$$
$$(\hat{\beta}_t)_i = \min\left\{C\sigma, \max\{c\sigma, (\tilde{\beta}_t)_i\}\right\}, \text{ if } (\tilde{\beta}_t)_i > 0, i \in \hat{S}_t,$$
$$(\hat{\beta}_t)_i = \min\{-c\sigma, \max\{-C\sigma, (\tilde{\beta}_t)_i\}\}, \text{ if } \tilde{\beta}_{t,i} < 0, i \in \hat{S}_t.$$

The details on the selection of parameters $\alpha_0, \eta_0, c, C, c_L$ will be postponed to Section IV.

B. Monitoring Statistics W_t

With the estimator $\hat{\beta}_t$, we are able to plug it into the CUSUM statistics in (5) for detection. For better illustration, we define the monitoring statistics W_t in two steps. Firstly, we define an initial statistics $\tilde{W}_t$ based on the classical CUSUM update for the detection of change point:

$$\tilde{W}_t = \max\left\{W_{t-1} + \log\frac{f_{X_t\hat{\beta}_t}(y_t)}{f_0(y_t)}, 0\right\}.$$

If $\tilde{W}_t \geq A$ for some threshold $A > 0$ that will be defined later, we directly set $W_t = \tilde{W}_t$. If $\tilde{W}_t < A$, we need to check the amount of data that are used for detection. For this purpose, we introduce a new tuning parameter $q = q(A)$. If the length of the time window $[M(t)+1, t-1]$ is too long (i.e., $t - M(t) - 1 \geq q$), we discard these data, re-set W_t back to 0 and restart the estimation. If $t - M(t) - 1 < q$, we still set $W_t = \tilde{W}_t$. To be mathematically rigorous, we define

$$W_t = \begin{cases} \tilde{W}_t, \text{ if } t - M(t) - 1 < q_A \text{ or } \tilde{W}_t \geq A, \\ 0, \text{ if } t - M(t) - 1 \geq q_A \text{ and } \tilde{W}_t < A. \end{cases} \tag{9}$$

C. Candidate change point $M(t)$

With the defined monitoring statistics W_t in (9), let us define the candidate change point $M(t+1)$ for the next time instant $t+1$. If $W_t > 0$, which indicates the possible existence of a change point, we keep the candidate change point as it is. If $W_t = 0$, we reset $M(t+1)$ back to the current time t. This is to say,

$$M(t+1) = \begin{cases} t, \text{ if } W_t = 0, \\ M(t), \text{ if } W_t > 0. \end{cases} \tag{10}$$

We raise a global alarm at the stopping time T_{IR}

$$T_{\mathrm{IR}} = \inf\{t > 0 : W_t \geq A\}, \tag{11}$$

Note that the threshold $A > 0$ is the same as A used for the definition of monitoring statistics, and is selected to satisfy the ARL to false alarm constraint in (4). Our proposed algorithm can be described as follows.

Algorithm 1 Implicit-Regularization-Based Quickest Detection in High-Dimensional Linear Model.

```
Choose suitable magnitude α, step size η, tuning parameters s, S, c_L, q and threshold A.
Initialize W_0 = 0, M(1) = 0
while t > 0 do
    Observe data (y_t, X_t).
    Obtain historical data y_{M(t)+1,t-1}, X_{M(t)+1,t-1}.
    Obtain the estimator β̂_t.
    Update the monitoring statistics W_t.
    if W_t ≥ A then
        Raise an alarm.
    else if W_t < A then
        Update the candidate change point M(t+1).
    end if
    t = t + 1
end while
```

IV. Theoretical Properties

In this section, we establish the theoretical guarantees for our proposed stopping time T_{IR} in (11). We begin by establishing a choice of the threshold A that guarantees that the ARL to false alarm constraint in (4) is met.

Lemma 4.1: For any constraint $\gamma > 1$, if the threshold $A \geq \log\gamma$, then we have that

$$\mathsf{E}_\infty[T_{\mathrm{IR}}] \geq \gamma. \tag{12}$$

We are now ready to analyze the detection delay relationship of T_{IR}. The main challenge here as compared to the analysis in standard linear regression model without change points is how to characterize the value of candidate change point $M(\nu+1)$. If $M(\nu+1) < \nu$, the estimator $\hat{\beta}_t$ for $t > \nu$ is biased before the reset of candidate change point due to the fact that the data in the time-window $[M(t)+1, t-1]$ is a mixture of both pre-change and post-change data. Under such scenario, a low percentage of pre-change data in the time-window is crucial to the accuracy of the estimation. Furthermore, the dependence between the value of $M(\nu+1)$ and the condition $T_{\mathrm{IR}} > \nu$ brings additional challenge to the theoretical analysis. To tackle this issue, we have the following lemma on $M(\nu+1)$.

Lemma 4.2: For any $n = 1, 2, \cdots$,

$$\mathsf{P}_\nu[\nu - M(\nu+1) \geq n \mid T_{\mathrm{IR}} > \nu] \leq c_1\exp(-c_2 n). \tag{13}$$

for some constants $c_1, c_2 > 0$.

Lemma 4.2 shows that $\nu - M(\nu+1)$ has an exponential tail bound, which means that with large probability, the candidate change point $M(\nu+1)$ will be close to the change point ν and ensures the low percentage of pre-change data during the estimation for $t > \nu$.

Define the following Kullback-Leibler divergence of the post-change versus the pre-change distribution:

$$I(\beta^*) = \mathsf{E}_{X_t}[\mathsf{E}_{y_t}[\log\frac{f_{X_t\beta^*}(y_t)}{f_0(y_t)}]],$$

where y_t, X_t are generated from model (1) with $\beta_t = \beta^*$. We have the following theorem on the detection delay of our proposed stopping time T_{IR}.

Theorem 4.3: If the tuning parameters for the implicit regularization algorithm $\alpha_0, \eta_0, c_L, c, C, s, q$ satisfy:

$$\alpha_0 \leq \min\{\frac{1}{5p^2}, 1, \frac{1}{2}\sqrt{\beta^*_{\min}}, \frac{\frac{2}{c_L}\sigma\sqrt{\frac{\log p}{m}}}{3(\beta^*_{\max}\sigma)^2}\}, \quad (14)$$

$$\eta_0 \leq \frac{1}{20\beta^*_{\max}},\ c_L \leq c_3, \quad (15)$$

$$|\beta^*_{\min}| \geq c\sigma,\ |\beta^*_{\max}| \leq C\sigma,\ s \geq s^*, \quad (16)$$

$$q = c_5 e^{c_4 A}, \quad (17)$$

for some constant $c_3 > 0, 0 < c_4 < 1/2, c_5 > 0$. the detection delay of our proposed algorithm is bounded by:

$$D[T_{\mathrm{IR}}] \leq \frac{A}{I(\beta^*)} + \frac{C_1 m}{I(\beta^*)}\sqrt{\mathsf{D}[T_{\mathrm{IR}}]} + \frac{C_2\sqrt{A/I(\beta^*)}}{I(\beta^*)} + C_3. \quad (18)$$

for any $m \geq C_4 \log p$ and some constant C_1, C_2, C_3, C_4 not related to m, γ, p.

To help better understand Theorem 4.3, here we add a few remarks.

1) By letting $A = \log\gamma$ in relationship (18), we conclude that the detection delay of our proposed algorithm T_{IR} mainly consists of two parts: (1) the standard delay term $\log\gamma/I(\beta^*)$ in change point analysis, (2) the additional term $O(m\sqrt{\mathsf{D}[T_{\mathrm{IR}}]}/I(\beta^*))$ which results from the information loss during the estimation of the unknown coefficient.
2) When the number of observations per time step $m \ll \log\gamma$, our proposed stopping time T_{IR} is first-order asymptotically optimal in the sense of minimizing detection delay for each and every β^* in the region defined in Theorem 4.3. This is because it asymptotically attains the lower bound of the CUSUM procedure in (6) that knows the true value of β^*. Unfortunately if $m \geq O(\log\gamma)$, this conclusion no longer holds as the additional term plays a non-negligible role in the delay.
3) Our proposed algorithm included seven tuning parameters. The two most important parameters are α_0, η_0 which determines the performance of implicit regularization algorithm when estimating the post-change regression coefficient β^*. The other five parameters c, C, s, q, c_L are introduced for theoretical analysis and has insignificant impact in the numerical studies. In our paper, we select $\alpha_0 = 0.001, \eta_0 = 0.1, c = 0, C = \infty, s = p, q = +\infty, c_L = 1$.
4) We would like to comment on the computational complexity of our algorithm. At each time t, the computational complexity of T_{IR} is $O(mpL_t)$, and this is comparable to that of the explicit regularization methods such as LASSO. In addition, the memory requirement of our proposed method is $mp(t-M(t)-1)$. We conjecture that $t-M(t)-1$ can be reduced to $O(1)$ by recursively estimating β^*, which will be investigated elsewhere.

V. Numerical Results

In this section, we conduct Monte Carlo simulation studies to validate our theoretical results. The dimension is set to $p = 1000$. Assume the pre-change coefficient $\beta_0 = 0$ and the post-change coefficient $\beta^* = [1, 1, 1, 1, 0, \cdots, 0]^\top$.

Two baseline methods are considered:

- T_{CUSUM} : $\hat{\beta}_t = \beta^*$ for all t, which is unrealistic in real-world application since it assumes the true post-change coefficients are given.
- $\tilde{T}$: $\hat{\beta}_t = \bar{b} = [b, \cdots, b]^\top$. Here b is chosen to be $b = 4/p$, so as to maximize the Kullback information number

$$\mathsf{E}_{\beta^*}[\log(f_{X_t\bar{b}}(y_t)/f_0(y_t))] = \frac{m}{2\sigma^2}\left(\|\beta^*\|_2^2 - \|\beta^* - \bar{b}\|_2^2\right)$$
$$= \frac{m}{2\sigma^2}\left(4 - 4(b-1)^2 - (p-4)b^2\right)$$

For the fairness of comparison, only the estimators $\hat{\beta}_t$ at each time t are different for $T_{\mathrm{IR}}, \tilde{T}$, while the construction of statistics and the detection policy remains the same.

The detailed settings are presented as follows:

- $m = 20, p = 200, 400, \sigma = 1$.
- $\gamma = 1000, 2000, 5000, 10000, 20000$.
- $\alpha_0 = 0.001, \eta_0 = 0.1$.
- $c = 0, C = +\infty, c_L = 1, s = p, q_A = \infty$.

γ	T_{IR}	T_{CUSUM}	$\tilde{T}$
1000	3.165±0.038	1.002±0.001	15.484±0.185
2000	3.201±0.039	1.003±0.001	16.503±0.189
5000	3.250±0.039	1.004±0.001	17.789±0.195
10,000	3.287±0.039	1.005±0.001	18.746±0.198
20,000	3.323±0.039	1.007±0.001	19.710±0.202

TABLE I
Detection Delay of $T, T_{\mathrm{CUSUM}}, \tilde{T}$ for $p = 200$

γ	T_{IR}	T_{CUSUM}	$\tilde{T}$
1000	4.089±0.049	1.002±0.001	26.765±0.273
2000	4.147±0.049	1.003±0.001	28.460±0.277
5000	4.204±0.049	1.005±0.001	30.675±0.282
10,000	4.248±0.049	1.006±0.001	32.321±0.285
20,000	4.289±0.049	1.008±0.001	34.028±0.289

TABLE II
Detection Delay of $T, T_{\mathrm{CUSUM}}, \tilde{T}$ for $p = 400$

For each γ and $T_{\mathrm{IR}}, T_{\mathrm{CUSUM}}, \tilde{T}$, we first use the bisection method to find suitable threshold A to attain the false alarm constraint, and then simulate the detection delay under the special scenario when $\nu = 0$. All results are based on 10,000 Monte Carlo simulations.

From Table I, T_{CUSUM} performs the best, which is consistent with the optimality of CUSUM when the true post-change parameters are completely specified. However, this is infeasible in practice. In addition, our proposed stopping time T_{IR} has a better detection delay performance than $\tilde{T}$ for all scenarios, which shows the effectiveness of the estimation of the unknown parameters.

VI. Technical Details

In this section, we provide the complete proof for Lemma 4.1 and a high-level sketch of the proof for Theorem 4.3. More technical details are attached in the supplementary material.

Proof of Lemma 4.1:

Denote by $f_\mu(x)$ the probability density function (pdf) of Gaussian distribution with mean μ and covariance matrix $\sigma^2 I$. Consider the statistics M_t :

$$M_t = \sum_{n=1}^{t} \prod_{r=n}^{t} \frac{f_{X_r\hat{\beta}_r}(y_r)}{f_0(y_r)}.$$

We state that $M_t \geq \exp(W_t)$. This is because

$$W_t \leq \max_{n=1,\cdots,t} \sum_{r=n}^{t} \log \frac{f_{X_r\hat{\beta}_r}(y_r)}{f_0(y_r)},$$

and thus

$$\begin{aligned} \exp(W_t) &\leq \max_{n=1,\cdots,t} \exp(\sum_{r=n}^{t} \log \frac{f_{X_r\hat{\beta}_r}(y_r)}{f_0(y_r)}), \\ &= \max_{n=1,\cdots,t} \prod_{r=n}^{t} \frac{f_{X_r\hat{\beta}_r}(y_r)}{f_0(y_r)} \\ &\leq \sum_{n=1}^{t} \prod_{r=n}^{t} \frac{f_{X_r\hat{\beta}_r}(y_r)}{f_0(y_r)} = M_t. \end{aligned}$$

We then consider the stopping time

$$T_M = \inf\{t > 0 : M_t \geq \exp(A)\},$$

and based on the fact that $M_t \geq \exp(W_t)$ for all t, it is then clear that

$$T_M \leq T_{\mathrm{IR}}.$$

It is worth mentioning that if we denote by $\mathscr{F}_t$ the filtration generated by the observed data sequence, i.e.

$$\mathscr{F}_t := \sigma(y_1, \cdots, y_t, X_1, \cdots, X_t),$$

then our proposed estimator $\hat{\beta}_t$ is $\mathscr{F}_{t-1}$-adapted. It is easily seen that $\{M_t - t\}$ is a martingale sequence under P_∞, with respect to the filtration $\{\mathscr{F}_t : t \in \mathbb{N}\}$. To see this, note that $\hat{\beta}_t$ is $\mathscr{F}_{t-1}$-adapted, we have that

$$\mathsf{E}_\infty[M_t - t|\mathscr{F}_{t-1}] = M_{t-1} + 1 - t = M_{t-1} - (t-1).$$

Applying the optional sampling theorem obtains that

$$\mathsf{E}_\infty[T_{\mathrm{IR}}] \geq \mathsf{E}_\infty[T_M] = \mathsf{E}_\infty[M_{T_M}] \geq \gamma.$$

□

Sketch of Proof for Theorem 4.3:

The key idea in proving detection delay relationship (18) is to decompose our proposed stopping time T_{IR} into a series of sequential test. For this purpose, let us start with the simpler scenario when change occurs at time $\nu = 0$. Under such case, all observed data $(y_t, X_t)_{\mathbb{Z}_+}$ are under the post-change scenario, in the sense that $y_t = X_t\beta^* + \epsilon_t$ for all t.

We define the following sequential test $\mathcal{T}$:

$$\mathcal{T} = \inf\{t > 0 \text{ or } t = q : S_t = \sum_{n=1}^{t} \log \frac{f_{X_n\hat{\beta}_n}(y_n)}{f_0(y_n)} \notin (0, A)\}, \tag{19}$$

where $\hat{\beta}_n = 0 \in \mathbb{R}^p$ for $n = 1$, and for general $n \geq 2$, $\hat{\beta}_n$ is obtained by first applying the implicit regularization algorithm to the data $(y_{1:n-1}, X_{1:n-1})$.

$$y_{1,n-1} = [y_1^\top, \cdots, y_{n-1}^\top]^\top,\ X_{1,n-1} = [X_1^\top, \cdots, X_{n-1}^\top]^\top,$$

and then truncating the estimator obtained by implicit regularization. It is easily seen that our proposed stopping time T_{IR} under $\nu = 0$ can be written as:

$$T_{\mathrm{IR}} = \mathcal{T}_1 + \cdots + \mathcal{T}_w,$$

where sequential tests $\{\mathcal{T}_i\}_{i=1,2,\cdots}$ have the same distribution as sequential test (19) and w is the first time when the statistics in the sequential test (19) crosses the upper threshold A. We then have:

$$\begin{aligned} \mathsf{E}_0[T_{\mathrm{IR}} \mid T_{\mathrm{IR}} > 0] &= \mathsf{E}_0[T_{\mathrm{IR}}] \\ &= \mathsf{E}_0[\mathcal{T}_1 + \cdots + \mathcal{T}_w] \\ &= \sum_{i=1}^{\infty} \mathsf{E}_0[\mathcal{T}_i]\mathsf{P}_0(w \geq i) \\ &= \frac{\mathsf{E}_0[\mathcal{T}]}{\mathsf{P}_0(S_{\mathcal{T}} \geq A)}. \end{aligned}$$

We conclude that to bound the detection delay of our proposed stopping time T_{IR} when change occurs at time $\nu = 0$, it suffices to study $\mathsf{E}_0[\mathcal{T}]$ and $\mathsf{P}_0(S_{\mathcal{T}} \geq A)$ under P_0.

For the scenario when change occurs at general change point $\nu \geq 1$, the proof becomes much more complicated. The main challenge is that the estimator after the change-time might be constructed based on a mixture of pre-change and post-change data. For this purpose, we consider the following sequential test:

$$\begin{aligned} \mathcal{T}_{(k)} = \inf\{t - k, t \geq k+1 : \tilde{S}_t = \sum_{n=k+1}^{t} \log \frac{f_{X_n\hat{\beta}_n,\sigma^2 I}(y_n)}{f_{0,\sigma^2 I}(y_n)} \\ \notin (-W_k, A - W_k) \text{ or } t - M(k+1) = q_A\}, \end{aligned} \tag{20}$$

where W_k is the value of detection statistics at change-time k and $M(k+1)$ is the candidate change time at time $k+1$. The detection delay for this general scenario can be bounded via the analysis of $\mathcal{T}_{(k)}$ and $\mathsf{E}_0[T_{\mathrm{IR}} \mid T_{\mathrm{IR}} > 0]$. Due to the page limit, the complete proof will be presented elsewhere.

ACKNOWLEDGEMENT

Q. Xu and Y. Mei were partially supported by NSF grant DMS-2015405, NIH grant 1R21AI157618-01A1, and the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002378. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Y. Yu is partially supported by EPSRC EP/V013432/1 and the Philip Leverhulme Prize.

REFERENCES

[1] S. Arora, N. Cohen, W. Hu, and Y. Luo, "Implicit regularization in deep matrix factorization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[2] S. L. Smith, B. Dherin, D. G. Barrett, and S. De, "On the origin of implicit regularization in stochastic gradient descent," *arXiv preprint arXiv:2101.12176*, 2021.

[3] Y. Cao, A. Thompson, M. Wang, and Y. Xie, "Sketching for sequential change-point detection," *EURASIP Journal on Advances in Signal Processing*, vol. 2019, pp. 1–22, 2019.

[4] J. Geng, B. Zhang, L. M. Huie, and L. Lai, "Online change-point detection of linear regression models," *IEEE Transactions on Signal Processing*, vol. 67, no. 12, pp. 3316–3329, 2019.

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[6] C. Zou and P. Qiu, "Multivariate statistical process control using lasso," *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1586–1596, 2009.

[7] G. Ciuperca, "Real time change-point detection in a model by adaptive lasso and cusum," *Journal de la Société Française de Statistique*, vol. 156, no. 4, pp. 113–132, 2015.

[8] M. Pollak, "Optimal detection of a change in distribution," *The Annals of Statistics*, pp. 206–227, 1985.

[9] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[10] P. Zhao, Y. Yang, and Q.-C. He, "High-dimensional linear regression via implicit regularization," *Biometrika*, vol. 109, no. 4, pp. 1033–1046, 2022.

[11] T. Vaskevicius, V. Kanade, and P. Rebeschini, "Implicit regularization for optimal sparse recovery," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

□