Change-Point Estimation in High Dimensional Regression Models

Bingwen Zhang, Jun Geng and Lifeng Lai

Abstract—We consider high dimensional nonhomogeneous linear regression models with $p \rightarrow 0$ or $p \gg n$, where $p$ is the number of features and $n$ is the number of observations. In the model considered, the underlying true regression coefficients undergo multiple changes. Our goal is to estimate the number and locations of these change-points and estimate sparse coefficients in each of the intervals between change-points. This paper develops an approach to solve multiple change-points estimation problem in high dimensional linear regression model based on sparse group Lasso (SGL). We analyze the performance of our approach and prove several consistency results. In particular, under certain assumptions and using a properly chosen regularization parameter, we show that the estimation errors of linear coefficients and change-point locations can be expressed as functions of $n$, $p$ and $s$, where $s$ is the sparse level of each coefficient. From these functions, we can understand how the estimation errors scale with system parameters and identify conditions on system parameters under which the estimation errors diminish. Furthermore, we show that the estimation of change-points is always overfitting, which eliminates the risk of missing true change-points, and the isolated estimated change-points between true change-points does not occur, which implies that the estimated change-points are clustered around the true change points. We further extend our studies to general linear models (GLM) and prove similar results. Numerical simulations are provided to illustrate the effectiveness of our approach.

Index Terms—Change-point estimation; Consistency; High Dimensional Regression; Sparse Group Lasso; Sparsity.

I. INTRODUCTION

In signal processing, one often encounters the problem of estimating parameters of models from given observations. Recently, we have witnessed a huge research interest in high dimensional models, in which the number of features or parameters to be estimated $p$ is on the same order of or even larger than the number of data points or observations $n$, i.e. $p \rightarrow 0$ or $p \gg n$ [2]–[14]. Under such high dimensional settings, many recent works focus on adding various penalty terms to cost functions used in the model estimation [15]–[18]. By adding penalty terms that penalize model complexities, one can avoid producing overfitting models and can produce more interpretable results in practice.

In most of the existing works, it is often assumed that observations are from one underlying model. This assumption is not appropriate for scenarios where the underlying model undergoes changes. For example, in building economic models, it is more desirable in practice to assume that the underlying economic growth model changes over time [19]. In the analysis of array-based comparative genomic hybridization (array-CGH) data, the models for different segment of DNA vary [20]. In the analysis of time dependent Gaussian graphical model, the edge structure varies [21]. In all above examples, the observations may come from different models and we need to consider more than one set of model parameters in the problem. Hence it is of interest to identify the change-points between different models and further build models for each interval between change-points.

There are some interesting existing works focusing on estimating change-points in low dimensional regression models. In low dimension regression models, $p$ is assumed to be either a constant or $p \rightarrow 0$ as $n$ increases. One approach to solve this problem in the low dimensional case is based on dynamic programming (DP) [22]–[25], which requires the information about the true number of change-points. There are also works using other methods. For example, [26], [27] discuss change-points detection under a Bayesian setup, which assumes that there is a prior distribution on the possible locations of the change-points. [28] discusses a method to partition observations into different subsets. Similar to [26], [27], the model in [28] assumes a prior probability of each partition. Furthermore, the algorithm proposed in [28] requires precise knowledge of the distribution of the observations and has a very high complexity. In [29], the authors propose a variant of Lasso to solve one dimensional (i.e., $p = 1$) multiple change-points estimation problem. In [30], [31], the authors propose to use group fused Lasso to solve multiple change-points estimation problem in low dimension. In our recent work [9], we propose a sparse group Lasso (SGL) based approach for change-points estimation in low dimensional linear regression models.

In this paper, we study multiple change-points estimation in the high dimensional case. In particular, we consider high dimensional regression problems in which the underlying true linear coefficients might undergo multiple changes and $p \rightarrow 0$. Our goal is to estimate the number and locations of change-points that divide available data into different regions, and further to produce sparse interpretable models for each region. We will employ the SGL based approach, proposed in our recent work [9] for the change-points estimation problem in the low dimensional case, for the high dimensional case. In the SGL based approach, through a linear transformation, we rewrite the model into an equivalent form. For this
equivalent form, the coefficients possess both inter and intra
group sparsity structure. Since inter and intra group sparsity
structure fit the scope of SGL, this change-points estimation
problem can be transformed into a SGL problem. In particular,
we reformulate the original linear regression with change-
points problem into a convex optimization problem with both
$\ell_1$ and $\ell_2$ penalties in the objective function. The solution
of this convex optimization problem then directly provides
the number and locations of change-points and the linear
coefficients of each region.

Although the SGL based approach for change-points es-
estimation was shown to possess desirable properties for low
dimensional models in [9], the analysis in [9] does not apply in
the high dimensional setting anymore, as it relies critically on
the assumption that $p$ is fixed as $n$ increases. In this paper, we
develop new tools to analyze the performance of the proposed
SGL based approach for multiple change-points estimation in
the high dimension setting. The overall strategy of our analysis
is to use contradiction. To be more specific, we focus on the
difference between the optimal value of the objective function
and the objective function evaluated at the true parameters of
the model. This difference should always be less than or equal
to zero due to the fact that the optimal solution achieves the
minimum of the objective function. Suppose some variables
satisfying some constraints can be the optimal solution, then
if that difference mentioned above is greater than zero, then
we form a contradiction. This contradiction means that those
constraints do not hold for the optimal solution. Then we
find properties of the optimal solution by reversing those
constrains. Using this strategy, under certain assumptions and
using a properly chosen regularization parameter, we show
that the estimation errors of linear coefficients and change-
point locations can be expressed as functions of the number
of observations $n$, the dimension of the model $p$ and the sparse
level of the model $s$. From the derived error functions, we can
characterize the conditions under which the proposed estimator
is consistent.

We further extend our study to general linear models
(GLM), which is a broader class of linear models and includes
classic models such as logistic regression models. We show
that using our approach, if the link function in GLM model
is strictly convex, then GLM enjoys the same consistency
properties as those of ordinary linear models except for some
constant scaling factors. The extension to GLM reveals a
broader area of potential applications of the proposed
approach.

In addition to the above mentioned work on the change-
point estimation in low dimensional models, our work is
also related to existing work on high dimensional uniform
models [18], [32]–[36]. [32] discusses the restricted eigenvalue
condition in Gaussian design matrices, which is quite useful
in high dimensional sparse models. [18], [32], [33] study high
dimensional estimation problems under uniform models. In
[33], the authors study high dimensional estimation under the
sparsity constraint that the parameters are in $\ell_s$ balls. In [18],
the authors show a very general approach to show that, under
the assumption that data are from one uniform model, one can
prove oracle consistency inequalities in the high dimensional

case. In [34], the authors study the change-points detection
problem in linear regression with identity design matrices.
[35], [36] consider the detection of change-point in high-
dimension data using low-dimension compressive measure-
ments in an online setting.

Our work is different from the works mentioned above in
several aspects. First, we consider nonhomogeneous models.
Second, we consider high dimensional setting. Third, we require
less information about the change-points. For example, we
do not need the number of change-points (as required
in the DP approach) nor the prior distribution of change-
points/partitions (as required in the Bayesian approach).

The remainder of the paper is organized as follows. In
Section II, we describe the model under consideration. In
Section III, we prove the consistency and properties of the
solution of our approach. In Section IV, we extend our study to
generalized linear models. In Section V, we provide numerical
examples to illustrate the performance of our approach. In
Section VI, we provide concluding remarks. In Appendix,
we provide proofs and supporting lemmas for the theoretical
results in this paper.

II. Model

A. Notation

Before presenting the details of our model, we first intro-
duce the notation convention used throughout the paper. We
use upper case boldface letters (e.g., $X$) to denote matrices
and lower case boldface letters (e.g., $x$) to denote column
vectors. For a matrix $X$, we use $X_{i,:}$ to denote the $i$th row
of $X$, and use $X_{:,j}$ to denote the $j$th column of $X$. For a
positive integer $k$, we use $[k]$ to denote $\{1, 2, \ldots, k\}$. We
define $[b, e] := \{b, b+1, \ldots, e\}$ where $b$ and $e$ are integers
with $e \geq b$. Similarly, $[b, e) := \{b, b+1, \ldots, e-1\}$. We use
$\mathbb{R}$ to denote the set of real number. Let $f$ be a function, we
use $\nabla f(x)$ to denote the gradient of $f$ at $x$. We use $c, c'$
and $c_1, c_2, \ldots$ to denote positive constants.

B. Problem Formulation

We consider the linear regression model

$$y_t = x_t^T \beta^*_t + \epsilon_t, \quad t \in [n],$$  \hspace{1cm} (1)

where $x_t \in \mathbb{R}^p$, $\beta^*_t \in \mathbb{R}^p$ is a sparse coefficients vector with
sparse level $s$, $n$ is the number of data points and $p$ is the
dimension of data, $\epsilon_t \in \mathbb{R}$ is the observation noise and $y_t \in \mathbb{R}$
is the response. We focus on a high-dimensional case in which
$p/n$ does not go to zero as $n \to \infty$. We assume that $\epsilon_t$'s are
independent and identically distributed (i.i.d.) with distribution
$\mathcal{N}(0, \sigma^2)$. Here $\mathcal{N}(0, \sigma^2)$ is the probability density function
(pdf) of Gaussian variables with zero mean and variance $\sigma^2$.

We focus on the scenario that the values of $\beta^*_t$'s change
over time. In particular, we assume that the linear model
experiences $K^*$ changes in the values of $\beta^*_t$'s. The value of
$K^*$ is unknown a priori. We denote the true change times by
t_{K^*}^k, k \in [K^*]$, with $t_0^* = 1$ and $t_{K^*}^k+1 = n+1$ by convention.
Hence, we have

$$\beta^*_t = \alpha^*_k, \quad \text{for } t \in [t_k^*-1, t_k^*),$$  \hspace{1cm} (2)
where \( \{\alpha^*_k, k \in [K^* + 1]\} \) denote the true values of coefficients, which are fixed but unknown. Throughout the paper, the superscript \((\cdot)^*\) denotes the true but unknown underlying value. Our goal is to estimate the number of change-points \(K^*\), the set of change-points \(\mathbb{T}^* := \{t^*_k, k \in [K^*]\}\), and the coefficients \(\{\alpha^*_k, k \in [K^*]\}\) through \(n\) pairs of observed data \((x_t, y_t), t \in [n]\). Fig. 1 illustrates the model under consideration.

![Fig. 1. Model](image)

To represent the model in a compact form, we define
\[
\begin{align*}
X & := \left( \begin{array}{ccc}
x_1^T & \cdots & x_n^T \\
\end{array} \right) \in \mathbb{R}^{n \times np}, \\
y & := (y_1 \cdots y_n)^T \in \mathbb{R}^n, \\
e & := (\epsilon_1 \cdots \epsilon_n)^T \in \mathbb{R}^n.
\end{align*}
\]
Let \(\beta^* \in \mathbb{R}^{np}\) denote the vector obtained by concatenating the true coefficients \(\beta^*_k\)’s for each data \((x_t, y_t)\). \(\beta^*\) consists of \(n\) subvectors, each of length \(p\).

With these notations, we can write the model (1) in a compact form:
\[
y = X\beta^* + e. \quad (3)
\]
In the following, we rewrite (3) in a slightly different form, which will provide motivation for our approach. Towards this end, we let \(\theta_t := \beta_t - \beta_{t-1}\) with \(\beta_0 := 0_{p \times 1}\). We note that for true coefficients \(\theta^*_t\), we have
\[
\theta^*_t = \begin{cases} 
\alpha^*_{t+1} - \alpha^*_k, & t = k^*_k, \ k = 0, \cdots, K^* \\
0_{p \times 1}, & \text{otherwise,}
\end{cases} \quad (4)
\]
in which \(\alpha^*_0 := 0_{p \times 1}\). In the following, we use \(\theta^* \in \mathbb{R}^{np}\) to denote the vector formed by concatenating each \(\theta^*_t, t = 1, \cdots, n\).

Define
\[
\hat{X} = \left( \begin{array}{ccc}
x_1^T & \cdots & x_k^T \\
x_2^T & \cdots & x_{k+1}^T \\
\end{array} \right) \cdots \left( \begin{array}{ccc}
x_n^T \\
\end{array} \right) \in \mathbb{R}^{n \times np}. \quad (5)
\]
The model in (3) can be written as
\[
y = \hat{X}\theta^* + e. \quad (6)
\]
Motivated by this unique sparsity structure of \(\theta^*\), we propose to solve the multiple change-points estimation problem via solving the following optimization problem
\[
\min_{\theta \in \mathbb{R}^{np}} \varphi(\theta) := \frac{1}{n} \|y - X\theta\|_2^2 + \lambda_n \left[ \gamma \sum_{t=1}^{n} \|\theta_t\|_2 + (1 - \gamma) \|\theta\|_1 \right], \quad (7)
\]
in which \(\mathcal{L}(\theta)\) is referred as the loss function, the function \(\mathcal{R}(\theta)\) is referred as the regularization penalty function, \(\lambda_n\) is the regularization penalty weight, and \(\gamma \in (0, 1)\) adjusts the relative weight between the two terms in \(\mathcal{R}(\theta)\). Here, the design of the penalty term \(\mathcal{R}(\theta)\) is motivated the unique sparsity structure of \(\theta^*\) discussed above. In particular, the term \(\sum_{t=1}^{n} \|\theta_t\|_2\) is used to encourage the group-wise sparsity in the solution, and the term \(\|\theta\|_1\) is used to encourage within group sparsity in the solution.

Problem formulation (7) is of the form of SGL [37]. This problem formulation is first proposed in our recent work [9] to solve multiple change-points estimation problem for the case of \(p\) being fixed while \(n\) increasing to infinity. In this paper, we consider the high-dimension case in which \(p\) also grows with \(n\). As the analysis in [9] relies crucially on \(p\) being fixed, the analysis approach used in [9] does not apply in this high-dimension case. In this paper, we will provide new analysis that works for the high-dimension case.

As there is a one-to-one correspondence between \(\theta\) and \(\beta\), the optimization problem in (7) can be equivalently written as
\[
\min_{\beta \in \mathbb{R}^{np}} \frac{1}{n} \|y - X\beta\|_2^2 + \lambda_n \gamma \sum_{t=1}^{n} \|\beta_t - \beta_{t-1}\|_2 \\
+ (1 - \gamma) \sum_{t=1}^{n} \|\beta_t\|_1. \quad (8)
\]
Let \(\hat{\theta}\) and \(\hat{\beta}\) denote the optimal solution of (7) and (8) respectively. \(\hat{\theta}\) and \(\hat{\beta}\) can be transformed to each other by noticing
\[
\hat{\theta}_t = \sum_{i=1}^{t} \hat{\theta}_i, \quad (9)
\]
\[
\hat{\beta}_t = \hat{\beta}_1 - \hat{\beta}_{t-1}. \quad (10)
\]
Furthermore, the locations of estimated change-points are those \(t\)’s such that \(\hat{\theta}_t\) is nonzero, and the estimated total number of change-points \(\hat{K}\) is the total number of nonzero \(\hat{\theta}_t\)’s.

Let
\[
\Delta_{\theta} := \hat{\theta} - \theta^*, \quad \Delta_{\beta} := \hat{\beta} - \beta^*,
\]
be the estimation error vectors. Directly from the definition above, we have
\[
X\Delta_{\beta} = \hat{X}\Delta_{\theta}. \quad (11)
\]
C. Assumptions on Data

Throughout the paper, we make following assumptions:
A1. \( x_t \) is generated from the distribution \( N(0, \Sigma) \), and \( 0 < l \leq \mu_{\min}(\Sigma) \leq \mu_{\max}(\Sigma) \leq L < \infty \), where \( l \) and \( L \) are constants independent of \( n \) and \( p \).

A2. \( \rho(\Sigma) = \max_{1 \leq i \leq p} \Sigma_{i,i} \leq \rho_c \), where \( \rho_c \) is a positive constant independent of \( n \) and \( p \).

A1 indicates that our design matrix is Gaussian ensemble as discussed in [18], [32], [33]. A2 puts a constraint on the covariance matrix of the Gaussian ensemble. Notice that the diagonal elements of \( \Sigma \) are variances of elements in \( x_t \). A2 means that the variance of each element of \( x_t \) is upper bounded by a constant.

We define following quantities with regards to the model
\[
I_{\min} := \min_{0 \leq k \leq K} |t_{k+1}^* - t_k^*|, \\
J_{\min} := \min_{0 \leq k \leq K} \|x_{\xi}^* - x_{\eta}^*\|_2.
\]

\( I_{\min} \) is the minimal interval between two consecutive true change-points. \( J_{\min} \) is the minimal \( l_2 \) distance between true coefficient vectors of two adjacent intervals.

In the next section, we will first derive general results without making particular assumptions on \( I_{\min} \) and \( J_{\min} \). We will then simplify the results when particular assumptions on these quantities are made. Those particular assumptions will be introduced in Section III-C.

III. Consistency

In this section, we develop consistency results of our approach and provide theoretical guarantees. We will introduce several useful notions and supporting lemmas before presenting the results.

A. Preliminary

We first recall the concept of dual norm and then present two lemmas that will be frequently used in the proof.

**Definition 1.** Let \( \Xi \) be a norm on \( \mathbb{R}^m \), its dual norm \( \Xi^* \) is given by
\[
\Xi^*(\theta) = \sup_{v \in \mathbb{R}^m \setminus \{0\}} \frac{\langle v, \theta \rangle}{\Xi(v)} = \sup_{\Xi(v) \leq 1} \langle v, \theta \rangle,
\]
in which \( \langle \cdot, \cdot \rangle \) denotes the inner product of two vectors.

**Lemma 1.** Let \( \tilde{\theta} \) be an optimal solution to the optimization problem (7) when we choose
\[
\lambda_n \geq 2R^*(\nabla L(\theta^*)) ,
\]
in which \( R^* \) is the dual norm of \( R \). Then we have
\[
R(\Delta_{\theta}) \leq 4R(\theta^*). \tag{12}
\]

**Proof:** Please see Appendix A-A.

**Lemma 2.** Let \( U \in \mathbb{R}^{m_1 \times m_2} \), suppose each row \( U_{i,:}, i = 1, \cdots, m_1 \) are independently generated using \( N(0, \Sigma_U) \) with \( 0 < l_U \leq \mu_{\min}(\Sigma_U) \leq \mu_{\max}(\Sigma_U) \leq L_U < \infty \), then we have
\[
\sqrt{l} \|v\|_2 \leq \|\Sigma_U^{1/2}v\|_2 \leq \sqrt{L} \|v\|_2. \tag{13}
\]

Furthermore, with a probability at least \( 1 - c' \exp(-cm_1) \), in which \( c, c' \) are universal positive constants from [32, Theorem 1], we have
\[
\frac{1}{m_1} \|Uv\|_2^2 \geq \frac{l}{32} \|v\|_2^2 - \frac{9}{4} \sqrt{L} \rho(\Sigma_U) \sqrt{\log m_2/m_1} \|v\|_2 \|v\|_1, \tag{14}
\]
hold for all \( v \in \mathbb{R}^{m_2} \) simultaneously.

**Proof:** Please see Appendix A-B.

B. Results for General Models

In this subsection, we present our main results regarding the estimation error for the general model specified in Section II. These results will be further simplified in Section III-C once we assume more details about the model.

Although the detailed proofs of the following propositions are long and tedious (as we need to properly address different cases), the basic idea is clear. In particular, let \( \hat{\theta} \) be the optimal value obtained from the optimization problem (7), then we have the following inequality
\[
\varphi(\hat{\theta}) \leq \varphi(\theta^*). \tag{15}
\]

The basic idea of the proofs is to show that, for (15) to hold, the bounds in the propositions must hold.

Before presenting the results, we define a quantity that will be frequently used
\[
\delta_n := \frac{64}{I_{\min}^2} \left( \frac{72}{(1-\gamma)^2} \sqrt{L} \rho(\Sigma) \sqrt{\frac{\log p}{n}} R^2(\theta^*) + 2 \lambda_n R(\theta^*) \right). \tag{16}
\]

In all results presented in this subsection, we assume that A1-A2 hold and we choose \( \lambda_n \) such that
\[
2R^*(\nabla L(\theta^*)) \leq \lambda_n < \frac{1}{2R(\theta^*)} \left( \frac{I_{\min} L^2_{\min}}{2n} \right) + \frac{72}{(1-\gamma)^2} \sqrt{L} \rho(\Sigma) \sqrt{\frac{\log p}{n}} R^2(\theta^*) . \tag{17}
\]

Before moving further, we would like to comment on (17), which puts lower and upper bounds on \( \lambda_n \). It is reasonable to have lower and upper bounds on \( \lambda_n \). If \( \lambda_n \) is too small, the solution of problem (7) will be similar to the solution of the problem without penalty and thus the solution will not possess the special sparsity structure. On the other hand, if \( \lambda_n \) is too large, the solution of problem (7) will not be able to capture change-points, as the estimated vector will become sparser as \( \lambda_n \) increases. The particular choice of lower bound stems from Lemma 1. The upper bound is equivalent to \( \frac{I_{\min}}{2n} > \delta_n \), which is a quantity that frequently arises in the detailed analysis. The particular form of upper bound comes from the application of (14) in the detailed analysis. It is natural to ask whether such \( \lambda_n \) exists and if so how to choose such \( \lambda_n \) without knowing the values of quantities such as \( \theta^* \) and \( I_{\min} \) etc. We will answer these questions in Section III-C which shows such \( \lambda_n \) exists under some mild assumptions and shows how to choose one.

Here we make an additional mild assumption on \( I_{\min} \) and
B1. $n \delta_n \to \infty$ as $n \to \infty$; $I_{\min} - 2n \delta_n \to \infty$ as $n \to \infty$.
This assumption can be satisfied for models in Section III-C. Furthermore, noticing that since we choose $\lambda_n$ satisfying (17), then we have $I_{\min} > 2n \delta_n$. So $I_{\min} \to \infty$ as $n \to \infty$.

Note that in assumption B1, we only require the quantities involved to increase, we do not put any requirement on the growth order.

Under assumptions A1, A2 and B1, we have the following propositions. Please refer to Appendix B for proof outline.

**Proposition 1.** If $\hat{K} = K^*$, then we have
\[
\frac{\max_{1 \leq k \leq K^*} |\hat{t}_k - t^*_k|}{n} \leq \delta_n. \tag{18}
\]
Proposition 1 shows that, if $\hat{K} = K^*$ and $\delta_n$ goes to zero as $n$ and $p$ increase, then the relative errors of estimated change-points locations diminish.

**Proposition 2.** If $\hat{K} = K^*$, then for $k \in [K^* + 1]$,
\[
\|\hat{\alpha}_k - \alpha^*_k\|^2 \leq \frac{32n}{(I_{\min} - 2n \delta_n)l} \left( \frac{36}{(1 - \gamma)^2} \sqrt{L \rho(\Sigma)} \sqrt{\log p \ n} R^2(\theta^*) \right) + 2\lambda_n R(\theta^*). \tag{19}
\]
The condition for Proposition 2 is the same as the conditions for Proposition 1. As will be shown in Section III-C, if more details of the model is provided, the result can be further simplified.

**Proposition 3.**
\[
\hat{K} \geq K^*. \tag{20}
\]
Proposition 3 shows that, if we cannot have $\hat{K} = K^*$, we can still guarantee that $\hat{K} \geq K^*$. If $\hat{K} < K^*$, then we miss some true change-points. Hence it is more desirable to have $\hat{K} = K^*$.

The results in Proposition 1 and 2 holds when $\hat{K} = K^*$ occurs. In the following, we show that even if $\hat{K} = K^*$ does not occur, we can still guarantee certain accuracy of the estimated change-points. Let $\hat{T}_K$ and $T^*$ be the set of estimated change-points and the set of true change-points respectively:
\[
\hat{T}_K := \{\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_K\}, \quad T^* := \{t_1^*, t_2^*, \ldots, t_K^*\}.
\]
For two sets $S_1$ and $S_2$, we define
\[
\varepsilon(S_1||S_2) = \sup_{s_2 \in S_2} \inf_{s_1 \in S_1} |s_1 - s_2|.
\]

Notice that $\max \{\varepsilon(S_1||S_2), \varepsilon(S_2||S_1)\}$ is the Hausdorff distance between $S_1$ and $S_2$ [38]. Using this notation, Proposition 1 can be restated as that $\varepsilon(\hat{T}_K||T^*) \leq n \delta_n$ and $\varepsilon(T^*||\hat{T}_K) \leq n \delta_n$ hold at the same time when $\hat{K} = K^*$. The following proposition is parallel to Proposition 1 for the case $\hat{K} > K^*$.

**Proposition 4.**
\[
\varepsilon(\hat{T}_K || T^*) \leq \delta_n.
\]
Proposition 4 implies that, for each true change-point, there is at least one estimated change-point that is less than $n \delta_n$ far away. However, this proposition does not exclude the case that some estimated change-points are far away from all true change-points, which indicates that this proposition does not exclude the existence of isolated estimated change-points between the interval of true change-points. The term isolated estimated change-point will be defined precisely in the sequel. Informally, an isolated estimated change-point is an estimated change-point that is far from the true change-points and other estimated change-points. In the following, we show that such events do not occur.

To proceed, we define event
\[
\mathcal{T}_{\text{isolate}} := \left\{ \exists j \in [t_k^*, t_{k+1}^* - 1] \text{ for some } k \in [K^* + 1] \right. \text{ s.t. } \min \{\hat{t}_{j+1}, t_{k+1}^*\} - \hat{t}_j \geq n \delta_n, \quad \hat{t}_j - \max \{\hat{t}_{j-1}, t_k^*\} \geq n \delta_n, \quad \|\hat{\alpha}_{j+1} - \hat{\alpha}_j\|_2 > J_{\min}, \right. \]
where $\delta_n$ is the nonnegative quantity defined in (16). The estimated change-points described in $\mathcal{T}_{\text{isolate}}$ are the isolated change-points with a big change in coefficients. The variable $\delta_n$ is used to measure the distance between the isolated change-point and its nearest true or estimated change-points. Figure 2 illustrates such an isolated change-point $\hat{t}_j$ for the case $\hat{t}_{j-1} \geq t_k^*$ and $\hat{t}_{j+1} \leq t_{k+1}^*$. The question arises naturally that whether such isolated estimated change-points can occur. The next proposition answers this question.

**Proposition 5.** $\mathcal{T}_{\text{isolate}}$ does not happen.

Proposition 5 shows that isolated change-points do not occur, and hence the estimated change-points are clustered around true change-points. As the result, we will have long ‘blank’ intervals, in which there are no estimated change-points, between true change-points. In the following, we show that, these long ‘blank’ intervals enable us to obtain a very good estimate of the linear coefficients.

To make the statements above precise, we first introduce some definitions. Let $\xi_k^i := [t_{k-1}^i, t_k^i]$ be the $k$th true interval. Let $\xi_i := [\hat{t}_{i-1}, \hat{t}_i]$ be the $i$th estimated interval, and $\hat{I}_i$ be the length of the $i$th estimate interval. We define the set $\mathcal{Y}(\varsigma_n) := \{ i \in [K + 1] | \xi_i \subset \xi_k^i \text{ for some } k \in [K^* + 1] \text{ and } \hat{I}_i > \varsigma_n \}$, where $\varsigma_n$ is some nonnegative quantity. In the following, we take $\varsigma_n = \frac{\min_{s_1, s_2} |s_1 - s_2|}{10}$. Then $\mathcal{Y}(\frac{\min_{s_1, s_2} |s_1 - s_2|}{10})$ contains all $i$’s such that the $i$th interval between estimated change-points is longer than $\frac{\min_{s_1, s_2} |s_1 - s_2|}{10}$. Here we pick $\varsigma_n = \frac{\min_{s_1, s_2} |s_1 - s_2|}{10}$ since we focus on long estimated intervals that are not too short compared to

![Fig. 2. Illustration of an isolated change-point $\hat{t}_j$ with $\hat{t}_{j-1} \geq t_k^*$, $\hat{t}_{j+1} \leq t_{k+1}^*$, $\hat{t}_{j-1} - \hat{t}_j \geq n \delta_n$, $\hat{t}_j - \hat{t}_{j-1} \geq n \delta_n$, and $\|\hat{\alpha}_{j+1} - \hat{\alpha}_j\|_2 > J_{\min}$.](attachment:image.png)
$I_{\text{min}}$. Actually, for our result to hold, $I_{\text{min}} = \frac{I_{\text{min}}}{c}$ would suffice, where $c > 0$. Let $\alpha_{k(i)}^*$ be the coefficients of “long” intervals containing $\xi_i$ for $i \in \Sigma_i$.

**Proposition 6** (Bounds for long intervals). For $i \in \Sigma \left( I_{\text{min}} \right)$, we have

$$
\|\hat{\alpha}_i - \alpha_{k(i)}^*\|_2^2 \leq \frac{320n}{I_{\text{min}}^2} \left( \frac{36}{(1 - \gamma)^2} \sqrt{L_0(\Sigma)} \sqrt{\log p} \right) \mathcal{R}^2(\theta^*) + 2\log n \mathcal{R}(\theta^*) .
$$

\[ (21) \]

C. Simplified Results with Knowledge of Model Details

The results provided in Section III-B are for general models. These results can be further simplified under more specific model assumptions. These simplified results can provide more insights. To proceed, we make the following additional assumption.

A3. $I_{\text{min}} = \Theta(n)$; $J_{\text{min}} = \Omega(1)$; $\mathcal{R}(\theta^*) = O(s)$; 

$$
\sqrt{\frac{\log p}{n}} \to 0 \text{ as } n \to \infty.
$$

Since $I_{\text{min}} \leq n$, then $I_{\text{min}} = \Theta(n)$ means that intervals between change-points are comparable with the total number of samples, i.e., the intervals between change-points are not too small. $J_{\text{min}} = \Omega(1)$ means the growth order of $J_{\text{min}}$ is at least the same as a constant, i.e. the “jump” between two consecutive intervals are not too small. Since there are at most $(K^* + 1)$ nonzero elements in $\theta^*$ and $K^*$ is a constant, we can assume $\mathcal{R}(\theta^*) = O(s)$ if we know each entry of the coefficients are bounded. $\sqrt{\frac{\log p}{n}} \to 0$ as $n \to \infty$ means that the growth order of the sparse level $s$ is not too big compared to the growth order of $(n/\log p)^{1/4}$.

In the following lemma, we show a particular choice of $\lambda_n$ has very desirable properties that will enable us to simplify the results obtained in Section III-B.

**Lemma 3.** If A1-A3 holds and we choose $\lambda_n = \frac{8\sigma_0\sqrt{dp}}{1 - \gamma} \log n \frac{1}{n}$, then the following hold

- $c_5 \frac{\log p}{n} \leq \delta_n \leq c_4 s^2 \sqrt{\frac{\log p}{n}}.$

- The right hand side of (19) can be upper bounded by

$$
\frac{c_4 s^2}{n} \sqrt{\frac{\log p}{n}}.
$$

- Condition (17) is satisfied with probability at least $1 - \frac{1}{n^p} - np \exp (-n)$.

**Proof:** See Appendix C.

Using Lemma 3, it is easy to see that assumption B1 is satisfied for large $n$ and we can rewrite Proposition 1-6 into simplified versions in the following remarks. In the following remarks, we assume A1-A3 hold, $n$ is large enough and we choose $\lambda_n = \frac{8\sigma_0\sqrt{dp}}{1 - \gamma} \log n \frac{1}{n}$.

**Remark 1.** If $\hat{K} = K^*$, then

$$
\max_{1 \leq k \leq K^*} \left| \hat{\alpha}_k - \alpha_k^\ast \right| \leq c_4 s^2 \sqrt{\frac{\log p}{n}},
$$

with probability at least $1 - \frac{1}{n^p} - np \exp (-n)$.

Note that a probability $1 - \frac{1}{n^p} - np \exp (-n)$ is introduced compared to Proposition 1. This is due to the choice of $\lambda_n = \frac{8\sigma_0\sqrt{dp}}{1 - \gamma} \log n \frac{1}{n}$, which satisfies condition (17) with a probability as shown in Lemma 3. Since we require (17) to be true for Propositions 1-6 to hold, hence in the following remarks, we add an extra probability term to each proposition.

**Remark 2.** If $\hat{K} = K^*$, then

$$
\max_{1 \leq k \leq K^*} \left| \hat{\alpha}_k - \alpha_k^\ast \right| \leq c_4 s^2 \sqrt{\frac{\log p}{n}},
$$

with probability at least $1 - \frac{2}{n^p} - np \exp (-n)$.

**Remark 3.**

$$
\hat{K} \gtrless K^*,
$$

with probability at least $1 - \frac{2}{n^p} - np \exp (-n)$.

**Remark 4.**

$$
\frac{c}{n} \left( \hat{T}_K || T^\ast \right) \leq c_4 s^2 \sqrt{\frac{\log p}{n}},
$$

with probability at least $1 - \frac{2}{n^p} - np \exp (-n)$.

**Remark 5.** $T_{\text{isolate}}$ does not happen with probability at least $1 - \frac{2}{n^p} - np \exp (-n)$.

These results illustrate how the estimation errors scale with the system parameters $p$, $n$ and $s$.

IV. General Linear Models

In this section, we extend our results obtained in linear models to general linear models (GLM) [18]. In GLM, $y_t$ and $x_t$ are related through the linear coefficient $\beta_t^*$ as

$$
P(y_t|x_t, \beta_t^*) \propto \exp \left\{ y_t (\beta_t^* x_t) - \Phi (\beta_t^* x_t) \right\}.
$$

Here $c$ is a fixed and known scale parameter. The link function $\Phi : \mathbb{R} \to \mathbb{R}$ is a strictly convex function. GLM model includes many important models as special cases. For example, if $c = \sigma^2$, $\Phi (h) = h^2/2$ and $y_t \in \mathbb{R}$, GLM covers the ordinary linear regression. If $c = 1$, $\Phi (h) = \log (1 + \exp (h))$ and $y_t \in \{0, 1\}$, it covers the logistic regression.

In the existing studies, e.g. in [18], it is assumed that there is no change in the underlying model, and hence $\beta_t^*$ is assumed to be a constant for all $t$. We will focus on GLM with change-points, i.e., we assume $\beta_t^*$ is not constant anymore. In particular, similar to Section II, we assume that the parameter $\beta_t^*$ is piece-wise constant with respect to $t$. More specifically, let $K^*$ be the number of changes in $\beta_t^*$, then for $k \in [K^*]$, 

$$
\beta_t^* = \alpha_k^*, \quad \text{for } t \in [t_k^\ast, t_{k+1}^\ast), \quad \text{and } \alpha_k^* \neq \alpha_{k-1}^*.
$$
where $t^*_0 := 1$ and $t^*_k := n + 1$ by convention. Following the same approach as in the previous sections, we let $\theta_t := \beta_t - \beta_{t-1}$, and let $\theta^* \in \mathbb{R}^{np}$ denote the vector formed by concatenating each $\theta^*_t$, $t = 1, \ldots, n$.

Again, since there are only $K^*$ change-points among $n$ observations, hence in $\theta^*$ only $K^* + 1$ subvectors are nonzero. Furthermore, as all subvectors are sparse vectors, $\theta^*$ is a $2(K^* + 1)\delta$-sparse vector indicating that $\theta^*$ has a unique sparse structure: $\theta^*$ has group-wise sparsity and sparsity within each group.

Using the same linear transformation and follow the same reasoning in Section II, we propose to perform the change-points estimation problem by solving

$$
\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{t=1}^{n} (-y_t \langle \theta, \tilde{x}_t \rangle + \Phi(\langle \theta, \tilde{x}_t \rangle)) \right\} + \lambda_n \left[ \gamma \sum_{t=1}^{n} \| \theta_t \|_2 + (1 - \gamma) \| \theta \|_1 \right] \right\}, \tag{22}
$$

where $\tilde{x}_t$ is the $t$th row of $\tilde{X}$ as defined in (5).

Similar to the linear model, from the optimal solution $\hat{\theta}$, we can then obtain the estimated set of change-points $\hat{T}_K := \{ \hat{t}_k, k \in [K] \}$, the estimated coefficients $\{ \hat{\alpha}_k, k \in [K + 1] \}$ of each region.

Similar to the previous section, we define

$$
\delta_n := \frac{64}{\mu J^2_{\text{min}}} \left( \mu \frac{72}{(1-\gamma)^2} \sqrt{L} \rho(\Sigma) \sqrt{\log p} / n \right) \mathcal{R}^2(\theta^*) + 6\lambda_n \mathcal{R}(\theta^*). \tag{23}
$$

In all results presented in this subsection, we assume that A1-A2 and B1 hold and we choose $\lambda_n$ such that

$$
2\mathcal{R}^* (\nabla \mathcal{L}(\theta^*)) \leq \lambda_n \leq \frac{1}{6\mathcal{R}(\theta^*)} \left( \frac{L_{\text{min}} \mu J^2_{\text{min}}}{64} \right) \mathcal{R}^2(\theta^*) - \mu \frac{72}{(1-\gamma)^2} \sqrt{L} \rho(\Sigma) \sqrt{\log p} / n \mathcal{R}^2(\theta^*), \tag{24}
$$

which implies that $\delta_n < \frac{l_{\text{min}}}{2\mu}$. Under these assumptions, the following propositions show that the results obtained for the linear regression case also hold for GLM. The proofs of these propositions can be found in Appendix D.

**Proposition 7.** If $\hat{K} = K^*$, we have $\max_{1 \leq k \leq K^*} \| \hat{\theta}_{k+1} \| \leq \delta_n$.

**Proposition 8.** If $\hat{K} = K^*$, then for $k \in [K + 1]$, $\| \hat{\alpha}_k - \alpha^*_k \|_2^2 \leq \frac{32n}{\mu (l_{\text{min}} - 2\delta_n)} \left( \mu \frac{72}{(1-\gamma)^2} \sqrt{L} \rho(\Sigma) \sqrt{\log p} / n \mathcal{R}^2(\theta^*) \right) + 6\lambda_n \mathcal{R}(\theta^*)$.

**Proposition 9.** $\hat{K} \geq K^*$.

**Proposition 10.** $\frac{\| \hat{\theta}_{K^*} \|_2}{\sqrt{n}} \leq \delta_n$.

**Proposition 11.** $T_{\text{isolate}}$ does not occur.

**Proposition 12 (Bounds for long intervals).** For $i \in \Gamma(l_{\text{min}}/10)$, $\| \hat{\alpha}_i - \alpha^*_i \|_2^2 \leq \frac{80n}{\mu l_{\text{min}}} \left( \mu \frac{72}{(1-\gamma)^2} \sqrt{L} \rho(\Sigma) \sqrt{\log p} / n \mathcal{R}^2(\theta^*) \right) + 6\lambda_n \mathcal{R}(\theta^*)$.

V. NUMERICAL SIMULATION

Our simulation is based our algorithm in [9], which has an R implementation in the package ChangePointCalc [39].

We first test our algorithm on synthesised data for the ordinary linear regression. We set $n = 800$, $p = 200$ and $s = 10$. In particular, we set the first $s$ coefficients to be nonzero in each $\beta_i$. In our simulation, we set $\gamma = 0.927$, $n = 800$, $K^* = 3$, and the real change-points are at 101, 301, 701. For each $\alpha^*_k$, $k = 1, \cdots, 4$, the first $s$ coefficients are drawn independently from uniform distribution $U[-2, 2]$ and others are set to zero. Each $x_{t,j} \sim \mathcal{N}(0, 1)$, and the noise $\epsilon_t \sim \mathcal{N}(0, 0.01)$. In Figure 3, x-axis represents the locations from 1 to $n$, and y-axis represents whether the data point at each location is an estimated change-point (1 means it is an estimated change-point, i.e., $\hat{\theta}_i \neq 0$, while 0 means $\hat{\theta}_i = 0$).

In Figure 4, y-axis represents the $l_2$-norm of each $\hat{\theta}_i$, $i \in [n]$. From the simulation results above, we can see that there is at least one estimated change-point around each true change-point. Furthermore, the change-points form clusters around true change-points. Hence it is easy to identify the number of change-points and the intervals between change-points from visualization results. These simulation results are consistent with our theoretical results obtained.

Next, we show our simulation results for logistic regression. The basic setting is the same as the simulation for the ordinary linear regression case. The only difference is that $n = 1000$ and the true change-points are at 401, 701. From Figures 5 and 6, we can again observe that there is at least one estimated change-point around each true change-point, and the change-points form clusters around true change-points. These figures confirm our conclusion that the proposed SGL based approach
is also effective in estimating multiple change-points for GLM.

VI. CONCLUSION

In this paper, we have studied a SGL based approach for estimating multiple change-points in high dimensional linear regression models. We have shown how to transform the high dimensional multiple change-points problem into a SGL based problem. We have shown that, under certain assumptions, the solution of the proposed approach enjoys desirable consistency properties. We have further extended our study to GLM and shown that the proposed approach also works for GLM.

APPENDIX A
PROOF OF SUPPORTING LEMMAS IN SECTION III-B

A. Proof of Lemma 1

Since \( \hat{\theta} \) is the optimal solution to the optimization problem (7), we have

\[
L(\hat{\theta}) + \lambda_n \mathcal{R}(\hat{\theta}) \leq L(\theta^*) + \lambda_n \mathcal{R}(\theta^*),
\]

which implies

\[
L(\hat{\theta}) - L(\theta^*) + \lambda_n \mathcal{R}(\hat{\theta}) \leq \lambda_n \mathcal{R}(\theta^*). \tag{25}
\]

For the \( L(\hat{\theta}) - L(\theta^*) \) term, we have

\[
L(\hat{\theta}) - L(\theta^*) = \frac{1}{n} \| y - \bar{X}(\theta^* + \Delta_\theta) \|^2 - \frac{1}{n} \| y - \bar{X} \theta^* \|^2
\]

\[
= \frac{1}{n} \Delta_\theta^T \bar{X}^T \bar{X} \Delta_\theta - \frac{1}{n} e^T \bar{X} \Delta_\theta
\]

\[
= \frac{1}{n} \Delta_\theta^T \bar{X}^T \bar{X} \Delta_\theta + \langle \nabla L(\theta^*), \Delta_\theta \rangle. \tag{26}
\]

In (26), the first term is related to the data and estimation error, the second term is also related to the noise vector \( e \). For the second term, if we choose \( \lambda_n \geq 2 R^* \langle \nabla L(\theta^*) \rangle \), we have

\[
\| \nabla L(\theta^*), \Delta_\theta \| \leq (a) R^* \| \nabla L(\theta^*) \| R(\Delta_\theta) \leq (b) \frac{\lambda_n}{2} R(\Delta_\theta). \tag{27}
\]

Plugging (27) in (26), we have

\[
\frac{1}{n} \Delta_\theta^T \bar{X}^T \bar{X} \Delta_\theta - \frac{\lambda_n}{2} R(\Delta_\theta)
\]

\[
\leq L(\hat{\theta}) - L(\theta^*) \leq \frac{1}{n} \Delta_\theta^T \bar{X}^T \bar{X} \Delta_\theta + \frac{\lambda_n}{2} R(\Delta_\theta). \tag{28}
\]

Using (28) and \( \frac{1}{n} \Delta_\theta^T \bar{X}^T \bar{X} \Delta_\theta \geq 0 \), we have

\[
L(\hat{\theta}) - L(\theta^*) \geq - \lambda_n \mathcal{R}(\Delta_\theta). \tag{29}
\]

Plugging (29) into (25), we have

\[
- \frac{\lambda_n}{2} \mathcal{R}(\Delta_\theta) + \lambda_n \mathcal{R}(\theta^* + \Delta_\theta) \leq \lambda_n \mathcal{R}(\theta^*),
\]

which coupled with the triangle inequality leads to

\[
- \frac{1}{2} \mathcal{R}(\Delta_\theta) + \mathcal{R}(\Delta_\theta) - \mathcal{R}(\theta^*) \leq \mathcal{R}(\theta^*).
\]

This is the same as

\[
\mathcal{R}(\Delta_\theta) \leq 4 \mathcal{R}(\theta^*).
\]
B. Proof of Lemma 2
If \( 0 < l_U \leq \mu_{\text{min}}(\Sigma_U) \leq \mu_{\text{max}}(\Sigma_U) \leq L_U < \infty \), we have
\[
\|\Sigma_U^{1/2}v\|_2^2 = v^T \Sigma_U^{1/2} \Sigma_U^{1/2} v \geq \mu_{\text{min}}(\Sigma_U)\|v\|_2^2 \geq l_U\|v\|_2^2,
\]
and
\[
\|\Sigma_U^{1/2}v\|_2^2 = v^T \Sigma_U^{1/2} \Sigma_U^{1/2} v \leq \mu_{\text{max}}(\Sigma_U)\|v\|_2 \leq L_U\|v\|_2^2,
\]
which completes the proof for (13).
From [32], for all \( v \in \mathbb{R}^{m_2} \) we have
\[
\frac{1}{\sqrt{m_1}}\|Uv\|_2 \geq \frac{1}{4}\|\Sigma_U^{1/2}v\|_2 - 9\rho(\Sigma_U)\sqrt{\frac{\log m_2}{m_1}}\|v\|_1
\]
with probability at least \( 1 - c' \exp(-cm_1) \).
Then using the inequality \( (a - b)^2 \geq \frac{1}{2}a^2 - ab \) along with (13), we obtain (14).

APPENDIX B
PROOF FOR CONSISTENCY RESULTS IN SECTION III-B
Here we provide the proof outline for proposition in Section III-B. We only provide key steps and proof structures due to limited paper length. Define
\[
F(\Delta_\theta):= \varphi(\theta^* + \Delta_\theta) - \varphi(\theta^*).
\]
For the optimal solution \( \theta \) to (7), we must have \( F(\Delta_\theta) \leq 0 \). The main idea of our proof is to analyze \( F(\Delta_\theta) \) to identify those \( \Delta_\theta \)'s such that \( F(\Delta_\theta) > 0 \). In particular, we show that if the conclusions in the propositions do not hold, we will have \( F(\Delta_\theta) > 0 \). In other words, for \( \theta \) to be the optimal solution to (7), the conclusions in the propositions must hold.

A. Proof for Proposition 1
We have
\[
F(\Delta_\theta) = \mathcal{L}(\theta) - \mathcal{L}(\theta^*) + \lambda_n \left( \mathcal{R}(\theta) - \mathcal{R}(\theta^*) \right)
\]
\[
 \geq \frac{1}{n} \left\| \mathbf{X} \Delta_\theta \right\|_2^2 - \frac{1}{2} \lambda_n \mathcal{R}(\Delta_\theta) + \lambda_n \left( \mathcal{R}(\theta) - \mathcal{R}(\theta^*) \right)
\]
\[
 \geq \frac{1}{n} \left\| \mathbf{X} \Delta_\theta \right\|_2^2 - \lambda_n \mathcal{R}(\Delta_\theta) + \lambda_n \left( \mathcal{R}(\theta) - \mathcal{R}(\theta^*) \right)
\]
\[
 \geq \frac{1}{n} \left\| \mathbf{X} \Delta_\theta \right\|_2^2 - 2\lambda_n \mathcal{R}(\theta^*), \tag{30}
\]
in which (a) comes from (28), and (b) is due to (11) and the triangle inequality.

Define events
\[
A_{n,j} := \left\{ |\hat{t}_j - t_j^*| < n\delta_n \right\},
\]
\[
C_n := \left\{ \max_{1 \leq k \leq K^*} |\hat{t}_k - t_k^*| < \min I_{\text{min}}/2 \right\}.
\]
Then we have
\[
1 - \sum_{j=1}^{K^*} \mathbb{P} \left\{ A_{n,j} \right\} \geq \mathbb{P} \left\{ C_n \right\}.
\]

To prove Proposition 1, it suffices to find upper bounds of the two probabilities \( \mathbb{P} \left\{ A_{n,j} \cap C_n \right\} \) and \( \mathbb{P} \left\{ A_{n,j} \cap C_n^c \right\} \) for each \( j \).

1) Bounding \( \mathbb{P} \left\{ A_{n,j} \cap C_n \right\} \): Define
\[
A_{n,j}^+: = \left\{ t_j^* - \hat{t}_j > n\delta_n \right\},
A_{n,j}^-: = \left\{ \hat{t}_j - t_j^* > n\delta_n \right\}.
\]
It suffices to prove our results under \( A_{n,j}^+ \cap C_n \) and the other case follows similarly. On \( A_{n,j}^+ \cap C_n \), we have \( t_k^* - \hat{t}_k < t_k^* - t_{k+1} \) for all \( k = 1, \ldots, K^* \). For \( j \) we have, \( t_j^* - \hat{t}_j > n\delta_n \) and \( n > \hat{t}_{j+1} - t_j^* > \frac{\log n}{2} \). Here we prove the case \( \hat{t}_{j+1} \leq t_j^* \) and the other case follows similarly. Now we turn to bound \( \mathbb{P} \left\{ A_{n,j}^+ \cap C_n \cap \{ \hat{t}_{j+1} \leq t_j^* \} \right\} \). The case under study is illustrated in Figure 7.

![Fig. 7. Illustration of the case \( \hat{t}_{j+1} \leq t_j^* \).](image)

using Lemma 2, we have
\[
\frac{1}{n} \left\| \mathbf{x}^T_{\hat{t}_j} \cdots \mathbf{x}^T_{t_j^*} \right\|_2^2 \leq \frac{n}{4} \sqrt{L_{\rho}(\Sigma)}
\]
\[
\geq \frac{1}{n} \left( \mathbf{x}^T_{\hat{t}_j} \cdots \mathbf{x}^T_{t_j^*} \right) \left( \mathbf{x}^T_{\hat{t}_j^*} - \hat{\alpha}_{j+1} \right) \right\|_2^2
\]
\[
= 1 - \sum_{j=1}^{K^*} \mathbb{P} \left\{ A_{n,j} \cap C_n \right\} - \sum_{j=1}^{K^*} \mathbb{P} \left\{ A_{n,j} \cap C_n^c \right\}.
\]

Fig. 7. Illustration of the case \( t_j^* - \hat{t}_j > n\delta_n, n > \hat{t}_{j+1} - t_j^* > \frac{\log n}{2} \), and \( \hat{t}_{j+1} \leq t_j^* \).

Here we use a proof technique that will be used repeatedly in later analysis. In this technique, we focus on one properly chosen interval for detailed analysis. In this case, we focus on the interval \( [\hat{t}_j, \hat{t}_{j+1} - 1] \). Continuing from (30), we have
\[
F(\Delta_\theta) \geq \frac{1}{n} \left\| \mathbf{x}^T_{\hat{t}_j} \cdots \mathbf{x}^T_{t_j^*} \right\|_2^2 - 2\lambda_n \mathcal{R}(\theta^*),
\]
\[
\geq \frac{1}{n} \left\| \mathbf{x}^T_{\hat{t}_j} \cdots \mathbf{x}^T_{t_j^*} \right\|_2^2 - 2\lambda_n \mathcal{R}(\theta^*),
\]
\[
= \frac{1}{n} \left\| \begin{pmatrix} \mathbf{x}^T_{\hat{t}_j} \\ \vdots \\ \mathbf{x}^T_{t_j^*} \end{pmatrix} \right\|_2^2 - 2\lambda_n \mathcal{R}(\theta^*),
\]
\[
+ \frac{1}{n} \left\| \begin{pmatrix} \mathbf{x}^T_{\hat{t}_j} \\ \vdots \\ \mathbf{x}^T_{t_j^*} \end{pmatrix} \right\|_2^2 - \alpha_{j+1} - \hat{\alpha}_{j+1} \right\|_2^2.
\]

Using Lemma 2, we have
\[
\frac{1}{n} \left\| \mathbf{x}^T_{\hat{t}_j} \cdots \mathbf{x}^T_{t_j^*} \right\|_2^2 \leq \frac{n}{4} \sqrt{L_{\rho}(\Sigma)}
\]
\[
\geq \frac{1}{n} \left( \mathbf{x}^T_{\hat{t}_j} \cdots \mathbf{x}^T_{t_j^*} \right) \left( \mathbf{x}^T_{\hat{t}_j^*} - \hat{\alpha}_{j+1} \right) \right\|_2^2
\]
\[
= 1 - \sum_{j=1}^{K^*} \mathbb{P} \left\{ A_{n,j} \cap C_n \right\} - \sum_{j=1}^{K^*} \mathbb{P} \left\{ A_{n,j} \cap C_n^c \right\}.
\]
\[
\frac{l}{32n} \| \alpha_j^* - \alpha_{j+1}^* \|_2^2 - \frac{9}{4} \sqrt{L} \rho(\Sigma) \\
\geq \frac{l}{32} \delta_n \| \alpha_j^* - \alpha_{j+1}^* \|_2^2 - \frac{9}{4} \sqrt{L} \rho(\Sigma)
\]
B. Proof for Proposition 2.6

The proof structures of Proposition 2.6 are quite similar. Hence we only show the proof for Proposition 2.2.

Since Proposition 1 holds, we have \(|t_k - t_k^*| \leq n\delta_n\) for each \(k = 1, \ldots, K^*\). Then if \(\|\alpha_k - \alpha_k^*\|_2^2 > \frac{32n}{(I_{\min} - 2n\delta_n)^l} \left( \frac{36}{(1-\gamma)^2} \sqrt{\mathbb{T}_p(\Sigma)} \frac{\log p}{n} R^2(\theta^*) + 2\lambda_n R(\theta^*) \right)\), we have

\[
\mathcal{F}(\Delta_\theta) \\
\geq \frac{1}{n} \|X\Delta_\theta\|^2 - 2\lambda_n R(\theta^*) \\
\geq \frac{1}{n} \sum_{t = t_{k-1}}^t (x_t^T (\hat{\beta}_t - \alpha_k^*)^2 - 2\lambda_n R(\theta^*)) \\
\geq \frac{1}{n} \sum_{t = t_{k-1} + 1} \sum_{n\delta_n} (x_t^T (\hat{\alpha}_t - \alpha_k^*)^2 - 2\lambda_n R(\theta^*)) \\
\geq \frac{m_k}{n} \frac{l}{32} \|\alpha_k - \alpha_k^*\|_2^2 - \frac{9}{4} \sqrt{\mathbb{T}_p(\Sigma)} \frac{\log p}{n} R^2(\theta^*) \\
\geq \frac{m_k}{n} \frac{l}{32} \|\alpha_k - \alpha_k^*\|_2^2 - \frac{9}{4} \sqrt{\mathbb{T}_p(\Sigma)} \frac{\log p}{n} R^2(\theta^*) \\
\geq \frac{1}{n} \left( I_{\min} - 2n\delta_n \right) \frac{1}{32} \|\alpha_k - \alpha_k^*\|_2^2 - \frac{36}{(1-\gamma)^2} \sqrt{\mathbb{T}_p(\Sigma)} \frac{\log p}{n} R^2(\theta^*) \\
\geq 0,
\]

in which (a) is due to (30), (b) is due to Lemma 2 with \(m_k = t_k - t_k^* - 2n\delta_n\) and holds with probability at least \(1 - c\exp(-c(I_{\min} - 2n\delta_n))\), and (c) is due to the definition of \(I_{\min}\), (34) combined with Lemma 1. Hence we have

\[
\mathbb{P}\left\{ \mathcal{F}(\Delta_\theta) > 0 \right\} \left| \|\alpha_k - \alpha_k^*\|_2^2 > \frac{32n}{(I_{\min} - 2n\delta_n)^l} \left( \frac{36}{(1-\gamma)^2} \sqrt{\mathbb{T}_p(\Sigma)} \frac{\log p}{n} R^2(\theta^*) + 2\lambda_n R(\theta^*) \right) \right| \\
\geq 1 - c\exp(-c(I_{\min} - 2n\delta_n)) > 0,
\]

where the last step is due to Assumption B1. Hence we have

\[
\|\alpha_k - \alpha_k^*\|_2^2 \leq \frac{32n}{(I_{\min} - 2n\delta_n)^l} \left( \frac{36}{(1-\gamma)^2} \sqrt{\mathbb{T}_p(\Sigma)} \frac{\log p}{n} R^2(\theta^*) + 2\lambda_n R(\theta^*) \right).
\]

Appendix C

Proof for Simplified Results in Section III-C

A. Proof of the first and second item in Lemma 3

First, we find an upper bound of \(\delta_n\). Under the additional assumption A3 and we choose \(\lambda_n = \frac{8\sigma\sqrt{n\rho}}{1-\gamma} \sqrt{\frac{\log np}{n}}\), using the definition of \(\delta_n\) in (16), we have that

\[
\delta_n \leq c_3 \left( s^2 \sqrt{\frac{\log p}{n}} + s \sqrt{\frac{\log np}{n}} \right),
\]

where (a) is due to \(J_{\min} = \Omega(1)\) and \(R(\theta^*) = O(s)\) in A3 and (b) is due to the fact that \(p/n \rightarrow 0\) as \(n\) increases.

Then we find a lower bound of \(\delta_n\). Note that in Assumption B1, we require \(n\delta_n\) to increase, so we need to give a lower bound of \(\delta_n\). Since \(R(\theta^*)/J_{\min} \geq 1\),

\[
\delta_n \geq \frac{64}{l} \left( \frac{72}{(1-\gamma)^2} \sqrt{\mathbb{T}_p(\Sigma)} \frac{\log p}{n} \left( \frac{R(\theta^*)}{J_{\min}} \right)^2 \right)
\]

\[
\geq c_5 \sqrt{\frac{\log p}{n}}.
\]

For the third item in Lemma 3, following the same reasoning for finding bounds for \(\delta_n\), we can show that the right hand side of (19) can be upper bounded by

\[
\mathbb{E}[\hat{\lambda}^2] \geq c_8 s^2 \frac{1}{n}.
\]

Thus we complete the proof for the first and second item in Lemma 3.

B. Proof of the third item in Lemma 3

In the following lemma, we first list several useful properties of the dual norm, whose proof can be found in [40].

Lemma 4. (a) Dual norm of \(\sum_{t=1}^n (\theta_t)^2\) is \(\max_{1 \leq t \leq n} \|\theta_t\|_2\).

(b) Dual norm of \(\ell_1\) is \(\ell_\infty\).

(c) Let \(\Xi_1\) and \(\Xi_2\) be two norms, if \(\Xi = \Xi_1 + \Xi_2\), then

\[
\Xi^*(\theta) = \min \max \Xi_1(\theta - z), \Xi_2(z).
\]

Let \(\alpha \in \mathbb{R}^+\) be a constant, then

\[
\Xi^*(\alpha\theta) = \Xi^*(\theta)/\alpha.
\]

Next, we compute an upper bound of the dual norm \(R^*\) of our penalty norm function \(R\), which will be used in later proof.

Lemma 5.

\[
R^*(\theta) \leq \|\theta\|_\infty/(1 - \gamma).
\]

Proof: Applying (47) and (48), we have

\[
R^*(\theta) = \min_{z \in \mathbb{R}^n} \left\{ \max_{1 \leq t \leq n} \|\theta - z\|_2/\gamma, \|z\|_\infty/(1 - \gamma) \right\} \\
\leq \|\theta\|_\infty/(1 - \gamma).
\]

Now, we state a lemma showing that by choosing \(\lambda_n = \frac{8\sigma\sqrt{n\rho}}{1-\gamma} \sqrt{\frac{\log np}{n}}\) we can satisfy the left side of (17) with a high probability.

Lemma 6. If A1 and A2 hold, and we choose \(\lambda_n \geq \frac{8\sigma\sqrt{n\rho}}{1-\gamma} \sqrt{\frac{\log np}{n}}\), then

\[
\lambda_n \geq 2R^*(\nabla\mathcal{L}(\theta^*))
\]

with probability at least \(1 - \frac{2}{np} n p \exp(-n)\).
Proof: First, we have
\[ \nabla L(\theta^*) = -\frac{2}{n} \tilde{X}^T e. \] (51)
Using Lemma 5 we have,
\[ 2R^*(\nabla L(\theta^*)) \leq \frac{4}{n(1-\gamma)} \|\tilde{X}^T e\|_{\infty}. \] (52)

To bound \( \|\tilde{X}^T e\|_{\infty} \), we first study the first \( p \) columns in \( \tilde{X} \) and other column follows. For \( 1 \leq j \leq p \), we have
\[ \Pr \left\{ \frac{1}{n} (\tilde{X}_{:,j})^T e \geq c \right\} \leq \exp \left( -\frac{nc^2}{10\sigma^2 \Sigma_{j,j}} \right), \]
where the step is due to the chi-square tail bound. Combining these two, we have
\[ \Pr \left\{ \frac{1}{n} (\tilde{X}_{:,j})^T e \geq c \right\} \leq 2 \exp \left( -\frac{nc^2}{10\sigma^2 \Sigma_{j,j}} \right) + \exp (-n). \]
As the result, we obtain
\[ \Pr \left\{ \left\| \frac{4}{n(1-\gamma)} \tilde{X}^T e \right\|_{\infty} \geq c \right\} \leq np \left( 2 \exp \left( -\frac{nc^2}{10\sigma^2 \Sigma_{j,j}} \right) + \exp (-n) \right), \]
which leads to
\[ \Pr \left\{ \left\| \frac{4}{n(1-\gamma)} \tilde{X}^T e \right\|_{\infty} \geq c \right\} \leq 2 \exp \left( -\frac{nc^2}{160\sigma^2 \Sigma_{j,j}/(1-\gamma)^2} + \log np \right) + \exp (-n + \log np). \]

Setting \( c^2 = \frac{320/(1-\gamma)^2 \sigma^2 \rho \log np}{n} \) and \( \lambda_n \geq \frac{8\sqrt{\rho} \log np}{1-\gamma} \sqrt{\frac{\log np}{n}} \), we obtain
\[ \Pr \left\{ \left\| \frac{4}{n(1-\gamma)} \tilde{X}^T e \right\|_{\infty} \geq \lambda_n \right\} \geq 1 - \frac{2}{np} - np \exp (-n). \]

Now, we show our choice of \( \lambda_n \) satisfies the right side of (17). Since \( s^2 \sqrt{\log p \over n} \to 0 \) in A3, then using (44), we know that \( \delta_n \to 0 \). Combining this with the assumption \( I_{\min} = \Theta(n) \) in A3, we have
\[ I_{\min}/2n \geq c_6 > \delta_n \to 0 \text{ as } n \to \infty, \]
and
\[ I_{\min}/2n - \delta_n \geq c_7. \] (54)
This means we can satisfy the condition \( I_{\min}/2n > \delta_n \), i.e.,
\[ \lambda_n \leq \frac{1}{2R^*(\nabla L(\theta^*))} \left( \frac{I_{\min}/2n - \delta_n}{2n} - \frac{72}{(1-\gamma)^2} \sqrt{\log p \over n} \right), \] (55)
which is the upper bound in (17).
Combining the result of Lemma 6 and (55), we have finished the proof for the third item in Lemma 3.

APPENDIX D
PROOF FOR RESULTS ON GLM IN SECTION IV
Here we follow the same proof outline as the Appendix B, and only show the key steps.
We first provide several supporting results that will be used throughout the proof.
First, as we choose
\[ \lambda_n \geq 2R^*(\nabla L(\theta^*)), \] (56)
then following similar steps as those in the proof of Lemma 1, we can prove
\[ \mathcal{R}(\Delta_\theta) \leq 4R^*(\theta^*). \] (57)
Second, motivated the proofs in Section III-B in which the main idea is to analyze two carefully chosen adjacent intervals, we show a property related two adjacent intervals that will be repeatedly used in the remainder of the proof. Second, motivated the proofs in Section III-B in which the main idea is to analyze two carefully chosen adjacent intervals, we show a property related to two adjacent intervals that will be repeatedly used in the remainder of the proof. For \( \hat{T}_K \) and \( T^\ast \) whose definitions are the same as those in Section III-B. Let \( \hat{T}_K \) and \( T^\ast \) whose definitions are the same as those in Section III-B. Let \( \hat{T}_K \) and \( T^\ast \) whose definitions are the same as those in Section III-B. Let \( \hat{T}_K \cup \hat{T}_K = \hat{T}_K \) and \( T^\ast \cup \hat{T}_K = \hat{T}_K \). In the following, we first show a property related to two adjacent intervals \( \hat{T}_j-1, \hat{T}_j \) and \( \hat{T}_j, \hat{T}_{j+1} \) as follows.

If \( \Phi \) is a strictly convex function, then \( \Phi'' > 0 \). Using Taylor expansion, we know that there exists \( \xi \in [x_i, x_i] \) such that
\[ \delta L(\Delta_\theta, \theta^*) = \Delta_\theta \left( \frac{1}{n} \sum_{t=1}^{n} (x_t^2 + \Phi'((\xi, x_i))) \right) \Delta_\theta \]
\[ \geq \frac{\mu}{n} \| X \Delta_\theta \|_2^2, \]
where $\mu = \min_{t \in \mathcal{E} \cap \mathcal{C}} \Phi'(\langle \xi_t, x_t \rangle) > 0$ and its value depends on the function $\Phi$, data matrix $X$ and true parameter $\theta^\star$.

Now, we focus on two adjacent intervals $[t_{j-1}, t_j]$ and $[t_{j}, t_{j+1}]$. Let $(\Delta_{t_{j-1}})$ denote the coefficients difference vector of the interval $t \in [t_{j-1}, t_j)$.

$$
\frac{\mu}{n} |X\Delta_{t_{j}}|^2 \\
\geq \frac{\mu}{32} \left( \frac{t_j - t_{j-1}}{n} \right) ||(\Delta_{t_{j-1}})||^2 + \frac{t_{j+1} - t_j}{n} ||(\Delta_{t_{j+1}})||^2 \\
- \mu \frac{72}{(1-\gamma)^2} \sqrt{L} \rho(\Sigma) \sqrt{\log p/n} R^2(\theta^\star),
$$

where we use Lemma 2, the fact $t_j - t_{j-1} \leq n$ and (57).

Then we use property (58) to study $\mathcal{F}(\Delta_{t_{j}})$.

$$
\mathcal{F}(\Delta_{t_{j}}) \\
\geq \frac{\mu}{32} \left( \frac{t_j - t_{j-1}}{n} \right) ||(\Delta_{t_{j-1}})||^2 + \frac{t_{j+1} - t_j}{n} ||(\Delta_{t_{j+1}})||^2 \\
- \mu \frac{72}{(1-\gamma)^2} \sqrt{L} \rho(\Sigma) \sqrt{\log p/n} R^2(\theta^\star) - 6\lambda_n R(\theta^\star),
$$

where we use the fact that we choose $\lambda_n \geq 2\mathcal{R}^\star(\nabla \mathcal{L}(\theta^\star))$ and (57), (58).

The proof strategy is to use (59) repeatedly as in Appendix B.

REFERENCES


