Change-points Estimation in Statistical Inference and Machine Learning Problems

by

Bingwen Zhang

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Doctor of Philosophy

in

Electrical and Computer Engineering

by

May 2017

APPROVED:

Professor Lifeng Lai, Major Thesis Advisor

Professor Donald R. Brown III, ECE Department, Worcester Polytechnic Institute

Professor Weiyu Xu, ECE Department, The University of Iowa

Abstract

Statistical inference plays an increasingly important role in science, finance and industry. Despite the extensive research and wide application of statistical inference, most of the efforts focus on uniform models. This thesis considers the statistical inference in models with abrupt changes instead. The task is to estimate change-points where the underlying models change.

We first study low dimensional linear regression problems for which the underlying model undergoes multiple changes. Our goal is to estimate the number and locations of change-points that segment available data into different regions, and further produce sparse and interpretable models for each region. To address challenges of the existing approaches and to produce interpretable models, we propose a sparse group Lasso (SGL) based approach for linear regression problems with change-points. Then we extend our method to high dimensional nonhomogeneous linear regression models. Under certain assumptions and using a properly chosen regularization parameter, we show several desirable properties of the method. We further extend our studies to generalized linear models (GLM) and prove similar results.

In practice, change-points inference usually involves high dimensional data, hence it is prone to tackle for distributed learning with feature partitioning data, which implies each machine in the cluster stores a part of the features. One bottleneck for distributed learning is communication. For this implementation concern, we design communication efficient algorithm for feature partitioning data sets to speed up not only change-points inference but also other classes of machine learning problem including Lasso, support vector machine (SVM) and logistic regression.

Acknowledgements

First and foremost, I shall greatly thank my research advisor, Dr. Lifeng Lai. He is not only a respectable and responsible person, but also provides valuable guidance, supports and excellent atmosphere for my research. It has been an honor to be his Ph.D. student. His enthusiasm for research is very encouraging for me when I am through the hard times in my research and Ph.D. study.

Thanks to my committee members, Dr. Donald R. Brown III and Dr. Weiyu Xu for sharing their time on this thesis. Dr. Brown's knowledge and guidance on signal estimation and detection laid a solid foundation for me in the area of signal processing and related research. Dr. Xu's innovative ideas on research helped me a lot on the way to do research during the time when I visited the UIowa. All of my committee members provide me guidance, supports and encouragement for my research and me. I appreciate each of them for their efforts and helps they provided during my Ph.D. study. The experience of working with these three scholars in my committee is a wealth of my life.

Thanks to all my lab mates Jun Geng, Ain Ul Aisha, Wenwen Zhao, Wenwen Tu and Mostafa El Gamal. They give me a lot of help both in life and on my research. Our lab has always been a source of friendships as well as good advice and collaboration. I am especially grateful for Jun Geng, the talks with him and his ideas helps me for my research.

I would like to thank my many friends and roommates during my time at WPI. They made the life more enjoyable and comfortable.

Thanks to my family. My mother Ziping Zhang and my father Zhenpeng Zhang, the most important persons for me, give me life, love and whatever I want unconditionally. A word of thanks A special word of thanks also goes to my family for their continuous support and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, we first introduce the background of statistical learning in homogeneous models. Then we list applications in which the model is not homogeneous and thus provide motivations for change-points inference in Section 1.1. In Section 1.2, we provide a summary of related works of this thesis. In Section 1.3, we list our main contributions of this thesis. In Section 1.4, we introduce common notations used throughout this thesis. Finally, in Section 1.5, we list the organization of this thesis.

1.1 Background and Motivation

1.1.1 Statistical Learning in Homogeneous Models

Statistical learning in homogeneous models, in which the data is assumed to be generated from a single underlying model, plays a key role in almost every branch in modern science and industry. Here, we give a few real life applications:

- Distinguish cancer versus normal patterns from mass-spectrometric data [1].
- Predict whether the income of a person exceeds $50K/year$ based on data feature

like person's age, work type, education and capital-gain etc [2].

- Identify relationship between murder rate and community population, per capita income, police operating budget and violent crime rate etc [3].
- Judge whether an email is a spam or not [2].

All these tasks are either linear regressions or classifications. Let n denote the number of sample, p denote the number of features. For example, for the third task mentioned above, murder rate is the target variable and thus we can use a vector $\mathbf{Y} \in \mathbb{R}^n$ to represent the murder rates of n given samples. All other variables such as community population, per capita income, police operating budget and violent crime rate can form a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, whose jth row $\mathbf{x}_j \in \mathbb{R}^p$ corresponds to jth sample. Thus the problem can be transformed into a linear regression and the task is to infer the linear relationship between $\mathbf{Y}$ and $\mathbf{X}$, which can be stated as

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \tag{1.1}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm. To avoid overfitting and to produce interpretable models,, usually an ℓ_1 penalty is added to (1.1) and thus forms Lasso [4]

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1, \tag{1.2}$$

where $\lambda > 0$ is the regularization parameter, and $\|\cdot\|_1$ denotes the ℓ_1 norm.

The other tasks mentioned above are binary classifications which can be done by either support vector machine (SVM) or logistic regression. Since they are binary classification tasks, we can take the positive labels as $+1$ and the negative labels as -1 and then we form a label vector $\mathbf{Y} \in \mathbb{R}^n$. And $\mathbf{X} \in \mathbb{R}^{n \times p}$ is formed the same way as linear regression.

Table 1.1: Examples of loss function and penalty function

Task	$\mathcal{L}(\boldsymbol{\beta})$	$\mathcal{R}(\boldsymbol{\beta})$
ordinary linear regression	$\frac{1}{2}\Vert\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\Vert^2$	0
Lasso	$\frac{1}{2}\Vert\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\Vert^2$	$\lambda\Vert\boldsymbol{\beta}\Vert_1$
SVM	$\sum_{j=1}^{n}\frac{1}{2}\max\left\{0,1-y_j\boldsymbol{\beta}^T\mathbf{x}_j\right\}$	$\frac{\lambda}{2}\Vert\boldsymbol{\beta}\Vert^2$
logistic regression	$\sum_{j=1}^{n}\log\left(1+\exp\left(-y_j\boldsymbol{\beta}^T\mathbf{x}_j\right)\right)$	0
ℓ_1 regularized logistic regression	$\sum_{j=1}^{n}\log\left(1+\exp\left(-y_j\boldsymbol{\beta}^T\mathbf{x}_j\right)\right)$	$\lambda\Vert\boldsymbol{\beta}\Vert_1$

For SVM, we solve the following optimization problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\sum_{j=1}^{n}\frac{1}{2}\max\left\{0,1-y_j\boldsymbol{\beta}^T\mathbf{x}_j\right\}+\frac{\lambda}{2}\Vert\boldsymbol{\beta}\Vert^2, \tag{1.3}$$

where y_j is the jth element of $\mathbf{Y}$. For logistic regression, we solve

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\sum_{j=1}^{n}\log\left(1+\exp\left(-y_j\boldsymbol{\beta}^T\mathbf{x}_j\right)\right). \tag{1.4}$$

Again to avoid overfitting and produce sparse results, we usually add a ℓ_1 norm penalty as in Lasso to have [5]

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\sum_{j=1}^{n}\log\left(1+\exp\left(-y_j\boldsymbol{\beta}^T\mathbf{x}_j\right)\right)+\lambda\Vert\boldsymbol{\beta}\Vert_1. \tag{1.5}$$

In these tasks above from (1.1) to (1.5), our goal is to infer a parameter vector $\boldsymbol{\beta}$ by minimizing a certain function. The optimization problem can be stated as

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\mathcal{L}(\boldsymbol{\beta})+\mathcal{R}(\boldsymbol{\beta}), \tag{1.6}$$

where $\mathcal{L}$ is *loss function* and $\mathcal{R}$ is *penalty function*. By adding penalty, the produced models have a certain sparsity structure, which is more interpretable and hence more desirable in practice. Lasso is such an example [4].

In Table 1.1, we list corresponding $\mathcal{L}(\boldsymbol{\beta})$ and $\mathcal{R}(\boldsymbol{\beta})$ in the above examples. We infer one parameter vector $\boldsymbol{\beta} \in \mathbb{R}^p$ by minimizing the sum of the loss function $\mathcal{L}(\boldsymbol{\beta})$ and penalty function $\mathcal{R}(\boldsymbol{\beta})$. This implies that $\boldsymbol{\beta}$ is same across all the samples. So one underlying assumption is that the models in the above tasks are homogeneous and all the data samples come from one uniform model, or the model is static. It is reasonable to make this assumption for the examples we mentioned at the beginning of this section. However, as will be discussed in the sequel, it might not be the case for some other applications.

1.1.2 Change-points Inference in Heterogeneous Models

As mentioned above, a typical assumption made in the existing work is that the data come from a single underlying model [4, 6–8]. However, this assumption might not hold in certain dynamic systems.

- In building economic growth models from various indicators, it is more appropriate to assume that the available data obeys different models over different time period as the economic growth pattern undergoes structural changes over the years [9].
- In the analysis of array-based comparative genomic hybridization (array-CGH) data, the underlying model varies in different segments of the DNA sequence [10].
- In the analysis of time dependent Gaussian graphical model, which has wide-spread applications in network traffic analysis and cyber attack detections, the edge structure varies [11].

In all above examples, it is of interest to identify the change-points and build proper models for different regions. This motivates the study change-points inference in statistical models.

For data that come from multiple underlying models, we cannot use homogenous models. Also, since the data are from different models, the learning algorithms in homogeneous models do not work. To address these issues, in this thesis we focus on learning algorithm design and analysis in heterogeneous models.

More specifically, we study the change-points inference problem in this thesis. The change-point identifies the shift point from one model to another. There are two typical formulations for change point problems: online and offline [12, 13]. In the online formulation, the observer receives observations sequentially. And the goal is to design real time algorithms to detect the change in the statistic behavior of the observations. To reduce the computational complexity, a statistic with a recursive form is desirable in the online detection. If such a recursive form exists, the statistic can be updated whenever a new observation arrives. For example, [14] proposes to track the gradual change of environmental parameters, and its statistics are updated recursively by minimizing a regret function.There are many other interesting papers focusing on the online detection problem, such as [15, 16]. In the offline formulation, initiated by [17], the observer is given a complete set of data and the goal is to estimate the location of change-points that segment the data set into several homogenous segments. The offline formulation has also attracted significant research interest (see survey [18] and Chapter 2.6 and 11 of [12]). Here, we list only a few of them to illustrate the its potential applications. For example, a direct application of the offline change-points estimation is data fitting [19]. The offline change-points estimation is also widely used in economic [20–22], molecular biology [10, 23], and climate data analysis [24].

In this thesis, we focus on the offline formulation with the goal of designing offline algorithms to estimate the location of change-points in a *given* data set. Since our data set is fixed, we do not focus on the recursive property of our algorithm. Instead, we mainly consider the consistency property and complexity of our estimator. This thesis mainly focuses on offline setting with both low dimensional and high dimensional cases studied. So we divide our work on offline change-points estimation in this thesis based on different data dimensions. Next, we introduce three main parts of this thesis on offline change-points estimation.

Thrust 1: Low Dimensional Change-points Inference

We begin by considering offline low dimensional linear regression problems in which the underlying true linear coefficients might undergo multiple changes. Our goal is to estimate the number and locations of change-points that segment available data into different regions, and further to produce sparse interpretable models for each region. The problem considered here has been studied extensively in other fields [12], and existing approaches to estimate multiple change-points are mainly based on least-square fitting via dynamic programming (DP) [25–28]. This DP approach will be discussed in detail in Chapter 2.

Although one can also apply the DP approach to solve this problem, there are several challenges associated with this approach. First, the DP algorithm cannot estimate the number of change-points accurately. It should be noticed that the DP approach needs information about the true number of change-points K^*. However, K^* in most cases are unknown. In particular, if we only know an upper-bound K_{max} on the total number of change-points, then the DP algorithm will always return K_{max} change-points. This is due to by adding new segments, one can always decrease the value of cost function. Hence, the DP algorithm cannot find the true number of change-points unless it is known perfectly. Second, the solution of the DP algorithm does not possess a sparse structure, hence, the model cannot be easily interpreted. Third, the computational complexity of the DP algorithm is high. In particular, for the model with K^* change-points, the computational complexity is $O(K^* n^3)$ with n being the total number of observations (samples).

To address these challenges, we propose to solve the change-points estimation problem using sparse group Lasso (SGL), a model fitting method proposed very recently in [29,30]. In SGL, the parameters are divided into groups. There are two penalty terms in the SGL problem formulation: the l_2 norm penalty, which encourages most of the groups of the solution to be zero, and the l_1 norm penalty, which will promote sparsity within those non-zero groups. We show that after a proper transformation, the parameters to be estimated possess both inter and intra group sparsity structure. Therefore, after a proper transformation, the problem studied in this thesis fits the scope of SGL and can be solved using SGL. In particular, we reformulate the original linear regression with change-points problem into a convex optimization problem with both l_1 and l_2 penalties. The solution of this convex optimization problem then directly provides the number and locations of change-points and the regression coefficients of each region. We prove that, under certain assumptions and a properly chosen regularization weight, the solution of the proposed approach possesses several desirable features: 1) the l_2 norm of the estimation errors of the linear coefficients diminishes as the number of available data increases; 2) the estimated locations of the change-points are close to those of the true change-points. We also propose a data-dependent method to choose a proper regularization weight. Furthermore, using efficient algorithms for solving SGL problems [30, 31], the complexity of the proposed approach is much lower than that of the DP approach.

Thrust 2: High Dimensional Change-points Inference

There is a growing interest in statistical inference 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. $\frac{p}{n} \nrightarrow 0$ or $p \gg n$ [32–34]. In this thrust, we focus on change-points estimation in high dimensional linear regression models.

Our goal remains the same as in that of the low dimensional setting. Although the SGL based approach for change-points estimation possesses desirable properties for low dimensional models, the analysis in the low dimensional setting does not apply in the high dimensional setting anymore, as it relies critically on the assumption that p is fixed as n increases. In the high dimensional setting, we develop new tools to analyze the performance of the proposed SGL based approach. 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 can 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.

Thrust 3: Speeding Up Inference Process

Here we consider how to solve change-points estimation problem more quickly. Since we have reformulated the change-points estimation problem as the SGL problem, so here we consider how to speed up solving SGL especially for high dimensional data. We are motivated by distributed learning techniques which are increasingly utilized due to the emergence of big data.

For big volume of data, the size of optimization problem is dramatically increasing and hence each machine cannot store all the data. The whole data set is split into parts and each part is stored in one machine. For a machine learning task in this scenario, each machine can only access its local data set and cannot access the whole data set. Since each machine can only store a part of data, the way to partition the whole data set is critical. There two popular ways to partition the data set: partitioning by sample [35] and partitioning by features [36, 37].

In SGL problem formulation, the features of the data are divided into groups. Hence we consider the way of data is storage is partitioning by features. In utilizing distributed learning for speeding up change-points estimation, one of the key steps is to communicate between nodes to update the parameters in the current optimization iteration. The amount of communication has become an bottleneck for speeding up distributed machine learning tasks [35–38]. This motivates us to design communication efficient distributed learning algorithms for feature partitioned data.

One major bottleneck in the design of large scale distributed machine learning algorithms is the communication cost. In this thrust, we propose and analyze a distributed learning scheme for reducing the amount of communication in distributed learning problems under the feature partition scenario. The motivating observation of our scheme is that, in the existing schemes for the feature partition scenario, large amount of data exchange is needed for calculating gradients. In our proposed scheme, instead of calculating the exact gradient at each iteration, we only calculate the exact gradient sporadically. In the iterations when exact gradients are not calculated, we will use the most recently calculated gradient as proxy to compute the next update. We provide precise conditions to determine when to perform the exact update, and characterize the convergence rate and bounds for total iterations and communication iterations. We further test our algorithm on synthesized and real data sets and show that the proposed scheme can substantially reduce the amount of data transferred between distributed nodes.

1.2 Related Efforts

For low dimensional setting, [14, 39–45], are most relevant to our work. [14] focuses on developing online algorithms to track a *gradually changing* parameter in the environment. Our work, on the other hand, focus on developing offline algorithms to estimate abrupt changes in a given data set. In [39], the authors proposed an adaption of Lasso algorithm to detect changes in the mean value of a sequence of Gaussian random variables and hence the dimension is one. In [40,41], the authors use group fused Lasso to solve the structural changes in linear regression problems. [42] considers the recovery of models that have multiple types of sparsity structure under a noiseless observation model. As will be clear in the sequel, in our work, two types of sparsity arises only in the transformed domain. This transformation imposes special constraints on the observation matrix, which does not satisfy the assumptions made in [42]. Furthermore, we consider noisy observation model and hence do not aim to recover the underlying signal exactly. [43, 44] discuss change-points detection under a Bayesian setup, i.e., there is a prior distribution on the possible locations of the change points, while this thesis is non-Bayesian. [45] discusses a method to partition observations into different subsets. Similar to [43, 44], the model assumes a prior probability of each partition. Furthermore, the algorithm needs precise knowledge of the distribution of the observations and has a very high complexity (exponential in n). Our work is different from these works in the following aspects. First, we impose an additional sparsity structure in the linear regression coefficient, which is often of interest in practice. Hence, instead of group fused Lasso, we use sparse group Lasso to solve the problem at the hand. The additional l_1 term in our problem formulation brings significant technical challenges when analyzing the performance of the algorithms. Moreover, we have analyzed the computational complexity of our proposed algorithm, while no such analysis was presented in [40, 41]. We also note that SGL has been used for anomaly detection in smart grid [46].

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 [8, 47–51]. [47] discusses the restricted eigenvalue condition in Gaussian design matrices, which is quite useful in high dimensional sparse models. [8, 47, 48] study high dimensional estimation problems under uniform models. In [48], the authors study high dimensional estimation under the sparsity constraint that the parameters are in ℓ_q balls. In [8], 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 [49], the authors study the change-points detection problem in linear regression with identity design matrices. [50, 51] consider the detection of change-point in high-dimension data using low-dimension compressive measurements 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).

For speeding up distributed learning, there have been a large number of recent interesting works on the sample partition scenario. For example, [52] and [53] proposed Communication Efficient Distributed Dual Coordinate Ascent (CoCoA) algorithm and its variant CoCoA+. In these algorithms, each machine solves a variant of a local dual problem and then updates the global parameters at each iteration. [54] designed Distributed Approximate Newton (DANE) algorithm. DANE is suitable for smooth and strongly convex problems and takes advantage of the fact that the subproblems at local machines are similar. In [36], Distributed Self-Concordant Optimization (DiSCO) algorithm was proposed. DiSCO uses an inexact damped Newton method and in each iteration step a distributed version of Preconditioned Conjugate Gradient (PCG) method is used to compute the inexact Newton step. In [55–59], variants of stochastic gradient descent (SGD) are proposed.

Compared with the sample partition scenario, the feature partition scenario is relatively less well understood. Among limited number of works on the feature partition scenario, in [35, 38], the authors propose to use randomized (block) coordinate descent to solve distributed learning problems for the feature partition scenario. In each iteration, each machine randomly picks a set of coordinates to optimize and apply updates to the parameters and gradients. As pointed out in [35] (will also be discussed in detail in the sequel), the communication cost associated with computing gradients, which are needed to calculate the next update, is very high.

Our work is also related but different from recent interesting work on the design of optimization algorithms with inexact updates. In [60], the convergence rate of inexact update of proximal method is proved. In [61], the optimal trade-off between convergence rate and inexactness is provided. In [62], the authors use an inexact method to solve distributed Model Predictive Control (MPC) problem. In [63], the authors analyze inexact updates in the coordinate descent. The main motivation for these works is to address the case that the subproblem for each machine cannot be solve exactly. In this thesis, however, we assume that each subproblem can be solved exactly and we try to introduce inexact updates or approximations to reduce communication cost. Our works is different from the works mentioned above in several aspects. First, this thesis focuses on feature partitioned data. Second, we use inexact updates to reduce communication cost, not based on the assumption that subproblem at each machine cannot be solved exactly. Third, we take a deterministic approach.

1.3 Contributions

We begin our work by studying multiple change-points estimation in low dimensional linear regression models. In this part, we list our contributions as follows [64].

- We transform the multiple change-points estimation problem into an SGL problem.
- We prove that the solution enjoys desirable properties: the estimation errors of the linear coefficients diminishes as the number of available data increases; the estimated locations of the change-points are close to those of the true change-points.
- We show that the complexity of the proposed approach is much lower than that of the DP approach.

Then we extend our results to high dimensional setting [65].

- 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.
- We can characterize the conditions under which the proposed estimator is consistent.
- We further extend our method to generalized linear models (GLM) and prove more general results.

For speeding up computation using distributed learning, we have the following contributions [66].

- We propose a communication efficient distributed learning algorithm to speed up a wide class of learning problems.
- We provide analytical results of communication amount of the proposed algorithm.
- We provide several practical techniques and simulations to show its feasibility in practice.

1.4 Notation

Here we introduce the notation convention used throughout this thesis. We use upper case boldface letters (e.g., $\mathbf{X}$) to denote matrices and lower case boldface letters (e.g., $\mathbf{x}$) to denote column vectors. For a matrix $\mathbf{X}$, we use $\mathbf{X}_{i,\cdot}$ to denote the ith row of $\mathbf{X}$, and use $\mathbf{X}_{\cdot,j}$ to denote the jth column of $\mathbf{X}$. For a positive integer k, we use $[k]$ to denote $\{1, 2, \cdots, k\}$. We define $[b, e] := \{b, b+1, \cdots, e\}$ where e and b are integers with $e \geq b$. Similarly, $[b, e) := \{b, b+1, \cdots, e-1\}$. We use $\mathbb{R}$ to denote the set of real number. Let f be a function, we use $\nabla f(\mathbf{x})$ to denote the gradient of f at $\mathbf{x}$. We use c, $c^{'}$ and $c_1, c_2, \cdots$ to denote positive constants.

1.5 Roadmap

Chapter 2 begins by introducing our problem formulation in low dimensional linear regression models. Then we show how to transform the multiple change-points estimation problem into an SGL problem. Theoretical guarantees are provided for the results of our approach. We provide simulation results of our approach.

Chapter 3 begins by extending our approach in low dimensional linear regression models to high dimensional linear regression models. Then we provide theoretical results for characterizing the estimation errors in expression of number of observations n, the dimension p and the sparse level s. Using these theoretical results, we can find the growth order of n, p and s to get a consistent estimator. We further extend our approach to GLM and prove corresponding analytical results, which implies our approach has a wide application range in practice.

Chapter 4 begins by introducing the concerns of high dimensional data which needs huge amount of computation power for change-points inference and other similar machine learning tasks. Hence we consider utilizing distributed learning. We propose a communication efficient distributed learning algorithm and propose analytical results for the amount of communication. Furthermore, we show how to set the parameters in the algorithm and show performance of our algorithm in practice.

Chapter 5 concludes the dissertation. It outlines the contributions of the dissertation and summarizes the thesis statement of this work.

Appendix A includes proof details and supporting lemmas.

Chapter 2

Low Dimensional Change-points Inference

This chapter begins the main part of this thesis by focusing on change-points inference in low dimensional linear regression models. In Section 2.1, we describe the model under consideration. In Section 2.2, we describe the proposed SGL based approach. In Section 2.3, we prove the consistency of the solution of our approach. In Section 2.4, the complexity of SGL algorithm is discussed. In Section 2.5, we provide numerical examples to validate the theoretic analysis.

2.1 Model

We consider the linear regression model

$$y_t = \boldsymbol{\beta}_t^{*T} \mathbf{x}_t + \epsilon_t, t \in [n], \tag{2.1}$$

where $\mathbf{x}_t$ is a p dimensional vector, $\boldsymbol{\beta}_t^*$ is a p dimensional sparse coefficients vector, where $p > 1$ is an integer, and ϵ_t is the observation noise. We assume that ϵ_t's are independent and identically distributed (i.i.d.) with $\mathcal{N}(0, \sigma^2)$. Here $\mathcal{N}(0, \sigma^2)$ is the probability density function (pdf) of Gaussian random variables with zero mean and variance σ^2.

We consider the scenario that the values of $\boldsymbol{\beta}_t^*$'s change over time. In particular, we assume that the linear model experiences K^* times of changes in the values of $\boldsymbol{\beta}_t^*$'s, and the set of change time instances (or change-points) are denoted as $T^* = \{t_k^*, k = 1, \ldots, K^*\}$. Hence, for $1 \leq k \leq K^* + 1$, we denote

$$\boldsymbol{\beta}_t^* = \boldsymbol{\alpha}_k^*, \quad \text{for } t_{k-1}^* \leq t \leq t_k^* - 1, \text{ and } \boldsymbol{\alpha}_k^* \neq \boldsymbol{\alpha}_{k-1}^*,$$

where $t_0^* = 1$ and $t_{K^*+1}^* = n + 1$ by convention, and $\{\boldsymbol{\alpha}_k^*, k = 1, \cdots, K^*\}$ are the true values of coefficients, which are fixed but unknown. Our goal is to estimate the change-points $\{t_k^*\}$, the coefficients $\{\boldsymbol{\alpha}_k^*\}$ and the number of change-points K^* through n pairs of observed data $(\mathbf{x}_t, y_t)$. Figure 2.1 illustrates the model.

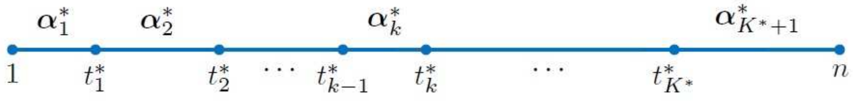
$\boldsymbol{\alpha}_1^*$
$\boldsymbol{\alpha}_2^*$
$\boldsymbol{\alpha}_k^*$
$\boldsymbol{\alpha}_{K^*+1}^*$
1
t_1^*
t_2^*
$\cdots$
t_{k-1}^*
t_k^*
$\cdots$
$t_{K^*}^*$
n

Figure 2.1: Illustration of model.

Let K_{max} be a known upper bound on the number of change-points and $K_{max} << n$, then the multiple change-points estimation problem can be written as

$$\begin{aligned} \min_{\boldsymbol{\beta}} \ & \frac{1}{n} \sum_{t=1}^{n} (y_t - \boldsymbol{\beta}_t^T \mathbf{x}_t)^2, \\ \text{s.t.} \ & \sum_{t=1}^{n-1} \mathbf{1}\{\boldsymbol{\beta}_{t+1} \neq \boldsymbol{\beta}_t\} \leq K_{max}, \end{aligned} \tag{2.2}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, whose value is 0 if $\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t$ and is 1 otherwise.

An intuitive approach to solve (2.2) is the exhaustive search, in which one solves a least square fitting problem for each possible change pattern, and picks the solution with the least residual square error. However, the total number of possible change patterns is $\sum_{K=0}^{K_{max}} \begin{pmatrix} n \\ K \end{pmatrix}$, which results in an extremely high computational complexity.

A more efficient way to solve (2.2) is to use DP as described below. Note that the theoretical analysis follows directly from discussion in [28], and the pseudocode is revised from DP algorithm in [67]. Let $\mathcal{M}_{K,n}$ be the set of all segmentations with K change-points ($K+1$ segmented intervals) up to nth sample. Let $r_k(m) = [t_k, t_{k+1})$ be the kth interval of segmentation m delimited by change-points t_k and t_{k+1}. Any segmentation m of $\mathcal{M}_{K,t}$ can be written as $\{[t_0, t_1), \cdots, [t_K, t_{K+1})\} = \{r_0(m), \cdots, r_K(m)\}$ with convention $t_0 = 1$ and $t_{K+1} = n + 1$. Then our task is to find an optimal segmentation $m \in \mathcal{M}_{K,n}$ to minimize to total cost. Our problem is to solve

$$\min_{m \in \mathcal{M}_{K,n}} \left\{ \sum_{r \in m} \min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \left\{ \sum_{i \in r} (y_i - \boldsymbol{\alpha}^T \mathbf{x}_i)^2 \right\} \right\}.$$

For any segment r, we define the cost as $g_r(\boldsymbol{\alpha}) = \sum_{i \in r} (y_i - \boldsymbol{\alpha}^T \mathbf{x}_i)^2$ and the optimal cost as $c_r = \min_{\boldsymbol{\alpha} \in \mathbb{R}^p} g_r(\boldsymbol{\alpha})$. Let $C_{K,n} = \min_{m \in \mathcal{M}_{K,n}} \left\{ \sum_{r \in m} c_r \right\}$. So we can retrieve the update equation

$$\forall t \geq K \; C_{K,t} = \min_{K-1 \leq j \leq t-1} \left\{ C_{K-1,j} + c_{[j+1,t]} \right\}. \tag{2.3}$$

CPE_DP solves multiple change-points estimation problem using DP, and PRINT_SOLUTION reconstructs and prints the solution.

In low dimensional case, the dimension p can be viewed as a constant. Hence the

Algorithm 1 CPE_DP($\mathbf{X}$,$\mathbf{Y}$,p,n,K^*)

```
let $r[0 \cdots K^*, 1 \cdots n]$ be a new matrix and $s[1 \cdots K^*, 1 \cdots n]$ be a new matrix
for $t = 1$ to $n$ do
    $r[0, t] = c_{[1,t]}$
end for
for $K = 1$ to $K^*$ do
    for $t = K$ to $n$ do
        $q = \infty$
        for $j = K - 1$ to $t - 1$ do
            if $q > r[K - 1, j] + c_{[j+1,t]}$ then
                $q = r[K - 1, j] + c_{[j+1,t]}$
                $s[K, t] = j + 1$
            end if
        end for
        $r[K, t] = q$
    end for
end for
return $r$ and $s$
```

Algorithm 2 PRINT_SOLUTION($\mathbf{X}$,$\mathbf{Y}$,p,n,K^*)

```
$(r, s)$ = CPE_DP($\mathbf{X}$,$\mathbf{Y}$,$p$,$n$,$K^*$)
$K = K^*$
while $K > 0$ do
    Print $s[K, n]$
    $n = s[K, n]$
    $K = K - 1$
end while
```

complexity to compute $c_{[j+1,t]}$ is $\Theta(t-j)$, then

$$\begin{aligned}
&\sum_{K=1}^{K^*}\sum_{t=K}^{n}\sum_{j=K-1}^{t-1}(t-j)\\
&=\sum_{K=1}^{K^*}\sum_{t=K}^{n}(t-K+2)(t-K+1)/2\\
&=\sum_{K=1}^{K^*}\left((n-K)(n-K+1)(2n-2K+1)/12+3(n-K)(n-K+1)/4+n-K\right).
\end{aligned}$$

From analysis above, we know the complexity of DP approach is $\Theta(K^* n^3)$[1].

Here we list two more drawbacks of this approach. First, the time complexity of DP approach $\Theta(K^* n^3)$ is still very high especially when n is large. Second, the solution of DP is not sparse in the sense that the estimated $\hat{\boldsymbol{\beta}}_t$s are not sparse vectors, which is not desirable when the interpretability of the model is important.

Motivating by the challenges of exhaustive search and DP approaches, we propose the SGL based approach for the proposed change-point estimation problem, which is described in the following section in detail.

2.2 Proposed SGL Based Approach

Let $\boldsymbol{\theta}_1^* = \boldsymbol{\beta}_1^*$ and $\boldsymbol{\theta}_t^* = \boldsymbol{\beta}_t^* - \boldsymbol{\beta}_{t-1}^*$ for $t = 2, \cdots, n$. Furthermore, let $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*T}, \boldsymbol{\beta}_2^{*T}, \cdots, \boldsymbol{\beta}_n^{*T})^T$, $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^{*T}, \boldsymbol{\theta}_2^{*T}, \cdots, \boldsymbol{\theta}_n^{*T})^T$. Notice that both $\boldsymbol{\beta}^*$ and $\boldsymbol{\theta}^*$ are np dimensional column vectors. From Section 2.1, we observe that most of $\boldsymbol{\theta}_t^*$ are zero vectors (there are at most K_{max} nonzero $\boldsymbol{\theta}_t^*$ vectors). Furthermore, for those non-zero $\boldsymbol{\theta}_t^*$'s, most of the entries in

[1] Throughout the thesis, $o_p(f(n)) = g(n)$ means that $\lim_{n\to\infty} P(|f(n)/g(n)| > \epsilon) = 0$ for any $\epsilon > 0$; $O_p(f(n)) = g(n)$ means that for any $\epsilon > 0$, there exists a finite $c > 0$ such that $P(|f(n)/g(n)| > c) < \epsilon$ for any n; $o(g(n)) = f(n)$ means that for any positive constant $c > 0$, there exists a constant $n_0 > 0$ such that $0 \le f(n) < cg(n)$ for all $n \ge n_0$; $O(g(n)) = f(n)$ means that there exist positive constants c and n_0 such that $0 \le f(n) \le cg(n)$ for all $n \ge n_0$; $\Theta(g(n)) = f(n)$ means that there exist positive constants c_1, c_2, and n_0 such that $0 \le c_1 g(n) \le f(n) \le c_2 g(n)$ for all $n \ge n_0$; $\Omega(g(n)) = f(n)$ means that there exist positive constants c and c_0 such that $0 \le cg(n) \le f(n)$ for all $n \ge n_0$.

$\boldsymbol{\theta}_t^*$ are zero, as $\boldsymbol{\beta}_t^*$'s are sparse vectors. As the result, if we view $\boldsymbol{\theta}_t^*$'s as groups within $\boldsymbol{\theta}^*$, then $\boldsymbol{\theta}^*$ has the following group sparse structure: most of the groups are zero, and for those non-zero groups, most of the entries within the group are zero.

Let $\mathbf{Y} = (y_1, y_2, \cdots, y_n)^T$, $\mathbf{e} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^T$ and

$$
\begin{aligned}
\mathbf{X} &= \begin{pmatrix} \mathbf{x}_1^T & & & \\ & \mathbf{x}_2^T & & \\ & & \cdots & \\ & & & \mathbf{x}_n^T \end{pmatrix}, \\
\tilde{\mathbf{A}} &= \begin{pmatrix} \mathbf{I}_p & & & \\ \mathbf{I}_p & \mathbf{I}_p & & \\ \cdots & \cdots & \cdots & \\ \mathbf{I}_p & \mathbf{I}_p & \cdots & \mathbf{I}_p \end{pmatrix},
\end{aligned}
$$

where $\mathbf{I}_p$ is the identity matrix of size $p \times p$. Hence $\mathbf{Y}$ and $\mathbf{E}$ are vectors of n dimension, and $\mathbf{X}$ and $\tilde{\mathbf{A}}$ are matrices of size $n \times np$ and $np \times np$, respectively. Define

$$
\tilde{\mathbf{X}} = \mathbf{X}\tilde{\mathbf{A}} = \begin{pmatrix} \mathbf{x}_1^T & & & & \\ \mathbf{x}_2^T & \mathbf{x}_2^T & & & \\ \mathbf{x}_3^T & \mathbf{x}_3^T & \mathbf{x}_3^T & & \\ \cdots & \cdots & \cdots & \cdots & \\ \mathbf{x}_n^T & \cdots & \cdots & \cdots & \mathbf{x}_n^T \end{pmatrix}, \qquad (2.4)
$$

then, it is easy to verify that our model can be rewritten as

$$
\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{e} = \tilde{\mathbf{X}}\boldsymbol{\theta}^* + \mathbf{e}. \qquad (2.5)
$$

To obtain the estimates of the number and locations of change-points and linear coefficients of each region, let $\boldsymbol{\beta}_0 = \mathbf{0}_{p\times 1}$, we propose to solve

$$\begin{aligned}\min_{\boldsymbol{\beta}} \quad & \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \gamma\lambda_n \sum_{t=1}^{n} \|\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}\|_2 \\ & +(1-\gamma)\lambda_n \sum_{t=1}^{n} \|\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}\|_1, \end{aligned} \tag{2.6}$$

which can also be rewritten as

$$\min_{\boldsymbol{\theta}} \frac{1}{n}\|\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\theta}\|_2^2 + \gamma\lambda_n \sum_{t=1}^{n} \|\boldsymbol{\theta}_t\|_2 + (1-\gamma)\lambda_n\|\boldsymbol{\theta}\|_1, \tag{2.7}$$

where $\|\cdot\|_2$ is the ℓ_2 norm, $\|\cdot\|_1$ is the ℓ_1 norm, λ_n is the regularization penalty weight, and $\gamma \in (0,1)$ adjusts the relative weight for the two penalty terms. γ affects the inter and intra group sparsity of the solution obtained from this optimization problem. The inter group sparsity is increased when we increase γ, while the intra group sparsity is increased when we decrease γ. Theoretically, as we can see from Proposition 1 and Proposition 2, as long as γ is a constant in $(0,1)$, we will have the consistent results in the change-point and coefficient estimations under proper assumptions. In practice, the choice of γ depends on the application. If one expects a strong group-wise sparsity, one should select γ to be a larger constant. If, in other applications, one expects mild group-wise sparsity, one should choose γ to be a smaller constant. We will discuss how to choose λ_n in Section 2.3. We note that the proposed problem formulation and algorithm do not depend on the parameter K_{max} or knowledge of K^*.

Notice that problem (2.7) is of the form of SGL proposed in [30]. As illustrated in [30], the penalty term $\sum_{t=1}^{n} \|\boldsymbol{\theta}_t\|_2$ encourages the group-wise sparsity, which implies that in the solution of (2.7) most of $\boldsymbol{\theta}_t$s are zero vectors, while $\|\boldsymbol{\theta}\|_1$ encourages sparsity within each group, which implies that in the solution of (2.7) most of entries are zero for those nonzero vectors. We also notice that (2.7) is a generalization of the problem considered in [41], in which a particular case with $\gamma = 1$ and $p = 1$ is considered.

Let $\{\hat{\boldsymbol{\theta}}_t\}$, $\{\hat{\boldsymbol{\beta}}_t\}$, $\hat{K}$ and $\hat{T}_{\hat{k}} = \{\hat{t}_k, k = 1, \cdots, \hat{K}\}$ denote estimates of $\{\boldsymbol{\theta}_t\}$, $\{\boldsymbol{\beta}_t\}$, K^* and T^*, respectively. For a given solution $\{\hat{\boldsymbol{\theta}}_t\}$ of (2.7), we can obtain $\{\hat{\boldsymbol{\beta}}_t\}$ from the linear relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. We can treat the nonzero vectors among $\hat{\boldsymbol{\theta}}_t$'s as change-points, from which the estimate of the total number $\hat{K}$ and locations $\hat{T}_{\hat{k}} = \{\hat{t}_k, k = 1, \cdots, \hat{K}\}$ of the change-points can be determined.

For $1 \leq k \leq \hat{K} + 1$, we denote

$$\hat{\boldsymbol{\alpha}}_k = \hat{\boldsymbol{\beta}}_t, \quad \text{for } \hat{t}_{k-1} \leq t \leq \hat{t}_k - 1, \text{ and } \hat{\boldsymbol{\alpha}}_k \neq \hat{\boldsymbol{\alpha}}_{k-1}.$$

By convention, we set $\hat{t}_0 = 1$ and $\hat{t}_{\hat{K}+1} = n + 1$.

2.3 Consistency

In this section, we study the properties of the solution of our SGL based approach (2.7). We provide consistency results and discuss how to choose the regularization parameter λ_n properly.

To assist the following presentation, we define

$$I_{min} = \min_{1 \leq k \leq K^*} |t^*_{k+1} - t^*_k|, \tag{2.8}$$

$$J_{min} = \min_{1 \leq k \leq K^*} \|\boldsymbol{\alpha}^*_{k+1} - \boldsymbol{\alpha}^*_k\|_2. \tag{2.9}$$

Hence, I_{min} is the minimal interval between two consecutive change-points, and J_{min} is the minimal ℓ_2 distance between two consecutive true different coefficient vectors.

Let $\{\delta_n\}$ be a sequence of positive quantities that decrease to zero as $n \to \infty$. Let $x_{t,m}$ denote the mth element of $\mathbf{x}_t$. Throughout this chapter, we make following assumptions:

A1:

$$
\begin{aligned}
0 &< l \le \inf_{\substack{1\le s<r\le n+1\\ r-s\ge n\delta_n}} \mu_{min}\left(\frac{1}{r-s}\sum_{t=s}^{r-1}\mathbf{x}_t\mathbf{x}_t^T\right)\\
&\le \sup_{\substack{1\le s<r\le n+1\\ r-s\ge n\delta_n}} \mu_{max}\left(\frac{1}{r-s}\sum_{t=s}^{r-1}\mathbf{x}_t\mathbf{x}_t^T\right) \le L < \infty,
\end{aligned}
$$

as $n \to \infty$, where $\mu_{min}(\cdot)$ and $\mu_{max}(\cdot)$ are the minimum and maximum eigenvalue of a matrix respectively.

Intuitively, A1 means that the eigenvalues of the averaged matrix $\left(\frac{1}{r-s}\sum_{t=s}^{r-1}\mathbf{x}_t\mathbf{x}_t^T\right)$ are bounded, which indicates that $\mathbf{x}_t\mathbf{x}_t^T$ is a well behaved matrix.

A2: $I_{min}/(n\delta_n) \to \infty,$ as $n \to \infty$.

A2 sets a requirement on the minimum intervals between any two consecutive change-points. In particular, we require I_{min} to grow as n grows. This assumption is reasonable as if I_{min} does not increase when n increases, then there exists an interval whose length is diminishingly small compared to n. It will be challenging to identify this interval from the whole data sequence.

A3: $\forall 1 \le m \le p$, $\forall 1 \le s < r \le n+1$ and $r - s \ge n\delta_n$,

$$
\frac{\ln n}{(r-s)^2}\sum_{t=s}^{r-1} x_{t,m}^2/J_{min}^2 \to 0.
$$

$\mathbf{x}_{t,m}^2$ can be viewed as the power of the mth dimension of $\mathbf{x}_t$. Intuitively speaking, A3 implies that J_{min}, the minimal ℓ_2 distance between two consecutive true different coefficient vectors, cannot be too small. This is a reasonable assumption, as if J_{min} is too small, there exists a change point at which the coefficient changes very little. It will be challenging to detect such as a change.

Proposition 1. *Under A1-A3, if $\hat{K} = K^*$, and we choose λ_n such that $\frac{\lambda_n}{J_{min}\delta_n} \to 0$ as*

$n \to \infty$, then

$$P\left(\max_{1 \le k \le K^*} \left|\hat{t}_k - t_k^*\right| \le n\delta_n\right) \to 1, \quad as \quad n \to \infty. \tag{2.10}$$

Proof. Please see Appendix A.2. □

Proposition 2. *Under A1-A3, if $\hat{K} = K^*$, and we choose λ_n such that $\frac{\lambda_n}{J_{min}\delta_n} \to 0$ as $n \to \infty$, then*

$$\|\boldsymbol{\alpha}_k^* - \hat{\boldsymbol{\alpha}}_k\|_2 \le \frac{n\lambda_n(\gamma + (1-\gamma)\sqrt{p}) + I_{min}o_P(J_{min})}{(I_{min} - 2n\delta_n)l}, \tag{2.11}$$

in probability as $n \to \infty$, for $\forall 1 \le k \le K^* + 1$.

Proof. Please see Appendix A.3. □

As discussed above, K^* is the true number of change-points, which is assumed to be a constant. Furthermore, from A2, we know that I_{min} is assumed to be sufficiently large, which implies that K^* cannot be arbitrarily large.

Remark 1. *Propositions 1 and 2 indicate that the proposed SGL based algorithm can lead to consistent estimations of the change-points and the linear coefficients. Proposition 1 is easy to interpret: from A2, we know that $n\delta_n/I_{min} \to 0$, which implies that the maximum relative change-points location estimate error is diminishing. Proposition 2 is a little complicated. Actually, there are several combinations of I_{min}, J_{min} that can make the estimation of the linear coefficients consistent. For example, if $I_{min} = \Theta(n)$ and J_{min} is a constant, then as long as $\lambda_n \to 0$, we have $\|\boldsymbol{\alpha}_k - \hat{\boldsymbol{\alpha}}_k\|_2 \to 0$. To see this, we first notice that $I_{min}o_P(J_{min})/((I_{min} - 2n\delta_n)l) \to 0$. Moreover,*

$$\frac{n\lambda_n(\gamma + (1-\gamma)\sqrt{p})}{(I_{min} - 2n\delta_n)l} = \frac{\lambda_n}{\delta_n}\frac{\gamma + (1-\gamma)\sqrt{p}}{(I_{min}/(n\delta_n) - 2)\,l} \to 0 \tag{2.12}$$

since $\lambda_n/\delta_n \to 0$ (from the assumption $\lambda_n/(J_{min}\delta_n) \to 0$ indicated in the condition of

Proposition 2 and J_{min} being a constant mentioned above) and $I_{min}/(n\delta_n) \to \infty$ as indicated in A2.

Remark 2. *Due to the special structure of $\tilde{\mathbf{X}}$ as shown in* (2.4)*, we cannot directly apply the existing bounds on the performance of regularized M-estimator (see, e.g., [8]). Hence we need a different approach to bound the l_2 norm of the error in the proof of Proposition 2.*

The above results require $\hat{K} = K^*$. In the following, we show that even if this assumption does not hold, we can still guarantee certain accuracy of the estimated change-points. For two sets $\mathbb{S}_1$ and $\mathbb{S}_2$, we define

$$\varepsilon(\mathbb{S}_1||\mathbb{S}_2) = \sup_{s_2 \in \mathbb{S}_2} \inf_{s_1 \in \mathbb{S}_1} |s_1 - s_2|. \tag{2.13}$$

Notice that $\max\{\varepsilon(\mathbb{S}_1||\mathbb{S}_2), \varepsilon(\mathbb{S}_2||\mathbb{S}_1)\}$ is the Hausdorff distance between $\mathbb{S}_1$ and $\mathbb{S}_2$ [68]. Since $\hat{\mathbb{T}}_{\hat{K}}$ and $\mathbb{T}^*$ are the set of estimated change-points and the set of true change-points respectively, so they can be written as

$$\begin{aligned} \hat{\mathbb{T}}_{\hat{K}} &:= \left\{\hat{t}_1, \hat{t}_2, \cdots, \hat{t}_{\hat{K}}\right\}, \\ \mathbb{T}^* &:= \left\{t_1^*, t_2^*, \cdots, t_{K^*}^*\right\}. \end{aligned}$$

Using this notation, Proposition 1 can be restated as that $\varepsilon(\hat{\mathbb{T}}_{\hat{K}}||\mathbb{T}^*) \le n\delta_n$ and $\varepsilon(\mathbb{T}^*||\hat{\mathbb{T}}_{\hat{K}}) \le n\delta_n$ hold at the same time in probability as $n \to \infty$ when $\hat{K} = K^*$. The following proposition is parallel to Proposition 1 for the case $K^* < \hat{K} < K_{max}$.

Proposition 3. *Under A1-A3, and choose λ_n such that $\frac{\lambda_n}{J_{min}\delta_n} \to 0$ as $n \to \infty$, then if $K^* < \hat{K} \le K_{max}$, we have*

$$P(\varepsilon(\hat{\mathbb{T}}_{\hat{K}}||\mathbb{T}^*) \le n\delta_n) \to 1, \quad as \quad n \to \infty. \tag{2.14}$$

Proof. Please see Appendix A.4. □

Proposition 3 implies that, if the number of the change-points is overestimated, i.e. $\hat{K} > K^*$, then there exists at least one estimated change-point falling in the range $n\delta_n$ of each true change-point. In the following, we show that the event $\{\hat{K} \geq K^*\}$ happens with a large probability.

We define

$$\begin{aligned} I_{max} &= \max_{1\leq k\leq K^*} |t^*_{k+1} - t^*_k|, \\ J_{max} &= \max_{1\leq k\leq K^*} \|\boldsymbol{\alpha}^*_{k+1} - \boldsymbol{\alpha}^*_k\|_2, \end{aligned}$$

and we impose another assumption:

A4. $I_{min} = \Theta(n)$; $J_{max} = O(1)$; $J_{min} = \Omega(1)$.

Proposition 4. *Under A1-A4, and we choose λ_n such that $\frac{\lambda_n}{J_{min}\delta_n} \to 0$ as $n \to \infty$, then*

$$P(\hat{K} \geq K^*) \to 1, \quad as \quad n \to \infty. \tag{2.15}$$

Proof. Please see Appendix A.5. □

Remark 3. *With A4, the conclusion of Proposition 2 can be further simplified. Notice that $J_{max} = O(1)$ indicates $o_P(J_{min}) \to 0$. Moreover, A2 and $\frac{\lambda_n}{J_{min}\delta_n} \to 0$ indicates $\frac{n\lambda_n}{I_{min}J_{min}} \to 0$, which further indicates $\frac{n\lambda_n}{I_{min}} \to 0$ as $J_{min} \leq J_{max}$. Therefore, we can conclude*

$$\frac{n\lambda_n(\gamma + (1-\gamma)\sqrt{p}) + I_{min}o_P(J_{min})}{(I_{min} - 2n\delta_n)l} \to 0.$$

That is, $\hat{K} = K^*$ with A1-A4 can guarantee the estimations of linear coefficients are consistent.

Remark 4. *From Proposition 4, we know that if A1-A4 are satisfied and λ_n is chosen* *such that $\frac{\lambda_n}{J_{min}\delta_n} \to 0$, we will have either: 1) $\hat{K} = K^*$, in which case we have Propositions 1 and 2 for the consistency of the estimates; or 2) $\hat{K} > K^*$, in which case, we have Proposition 3 for the consistency of the estimates. However, if $\hat{K} < K^*$, then some change-points are not detected. Hence it is more desirable to have $\hat{K} \geq K^*$.*

If one insists on having $\hat{K} = K^*$, we have the following data-dependent method to choose λ_n. This approach is based on the Akaike information criterion (AIC) [69]. For any given λ_n, we first solve (2.7) and obtain $\hat{K}$ and $\hat{T}_{\hat{K}} = \{\hat{t}_k, k = 1, \cdots, \hat{K}\}$ that divides the data into $\hat{K} + 1$ regions. We define

$$B(\hat{T}_{\hat{K}}) = \frac{1}{n} \sum_{k=1}^{\hat{K}+1} \sum_{t=\hat{t}_{k-1}}^{\hat{t}_k - 1} (y_t - \hat{\boldsymbol{\alpha}}_k^T \mathbf{x}_t)^2, \tag{2.16}$$

where $\hat{\boldsymbol{\alpha}}_k$ is the ordinary least squares (OLS) estimator in the interval $[t_{k-1}, t_k - 1]$.

Then we propose to minimize the cost function

$$C(\lambda_n) = \ln(B(\hat{T}_{\hat{K}})) + \rho_n p(\hat{K} + 1), \tag{2.17}$$

where ρ_n is a designed parameter such that $\rho_n \to 0$ and $\frac{\rho_n}{\delta_n} \to \infty$ as $n \to \infty$. In (2.17), $\ln(B(\hat{T}_{\hat{K}}))$ measures the accuracy of how well the model is fitted, and $\rho_n p(\hat{K} + 1)$ is the penalty of the number of estimated change-points.

Denote $\Omega = [0, \lambda_{max}]$, in which $\lambda_{\max}$ is the maximum value of λ_n such that the solution to (2.7) is not all zero vectors. $\lambda_{\max}$ can be easily computed. Define

$$\Omega_- = \{\lambda_n \in \Omega | \hat{K} < K^*\}, \Omega_+ = \{\lambda_n \in \Omega | \hat{K} > K^*\},$$

and λ^* is any λ_n such that $\hat{K} = K^*$.

Then, we have following proposition.

Proposition 5. *Under A1-A4, and $\frac{\lambda_n}{J_{min}\delta_n} \to 0$, we have*

$$P\left(\inf_{\lambda_n \in \Omega_+ \cup \Omega_-} C(\lambda_n) > C(\lambda^*)\right) \to 1, \quad as \quad n \to \infty.$$

Proof. Please see Appendix A.6. □

Remark 5. *Proposition 5 provides a method to choose the regularization parameter λ_n to guarantee the stronger result. In particular, if we choose $\lambda_n = \lambda^*$ and use this value in* (2.7)*, then we have $\hat{K} = K^*$ with probability* 1*. On the other hand, if one does not insist in having $\hat{K} = K^*$, λ_n can be simply chosen to satisfy $\lambda_n/(J_{min}\delta_n) \to 0$. Any value satisfying this condition will guarantee that the consistent results in Propositions 1 $\sim$ 4 hold.*

Remark 6. *If we set $p = 1$, the problem considered in this thesis becomes a basic Total Variance (TV)-regularization problem [70–72]. By setting $p = 1$ in the propositions above, we have the consistent results for this special case. Note that these results do not mean that we obtain full understanding of the TV-regularization problem. They simply imply that for the basic TV-regularization problem, we have certain consistency results regarding the change-points estimations and coefficient estimations under the assumptions made in this thesis.*

2.4 Complexity

In this section, we study the computational complexity of SGL algorithms. Denote the cost function in (2.7) as

$$\phi(\boldsymbol{\theta}) = \frac{1}{n}\left\|\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\theta}\right\|_2^2 + \gamma\lambda_n \sum_{t=1}^{n} \|\boldsymbol{\theta}_i\|_2 + (1-\gamma)\lambda_n \sum_{t=1}^{n} \|\boldsymbol{\theta}_i\|_1 , \tag{2.18}$$

and let ϕ^* be the minimum value of $\phi(\boldsymbol{\theta})$.

In our model, we have an $n \times 1$ output vector $\mathbf{Y}$, an $n \times np$ data matrix $\tilde{\mathbf{X}}$ which can be divided into n sub-matrices, $\tilde{\mathbf{X}}^{(1)}, \cdots, \tilde{\mathbf{X}}^{(n)}$, and each $\mathbf{X}^{(t)}$ is an $n \times p$ matrix for $t = 1, \cdots, n$, and an $np \times 1$ coefficient vector $\boldsymbol{\theta}$ which can be divided into n $p \times 1$ sub-vectors, $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_n$. The cost function (2.18) can be rewritten as

$$\begin{aligned}\phi(\boldsymbol{\theta}) &= \frac{1}{n}\left\|\mathbf{Y} - \sum_{t=1}^{n} \tilde{\mathbf{X}}^{(i)}\boldsymbol{\theta}_i\right\|_2^2 \\ &\quad + \gamma\lambda_n \sum_{t=1}^{n} \|\boldsymbol{\theta}_i\|_2 + (1-\gamma)\lambda_n \sum_{t=1}^{n} \|\boldsymbol{\theta}_i\|_1, \qquad (2.19)\end{aligned}$$

and $\tilde{\mathbf{X}}^{(i)} = \begin{pmatrix} \mathbf{0}_{p\times 1} & \cdots & \mathbf{0}_{p\times 1} & \mathbf{x}_i & \cdots & \mathbf{x}_n \end{pmatrix}^T$.

We define another function which will be used in further analysis. Define

$$\begin{aligned}\phi_{inner}(\boldsymbol{\theta}_k) &= \frac{1}{n}\left\|r(-k) - \tilde{\mathbf{X}}^{(k)}\boldsymbol{\theta}_k\right\|_2^2 + \gamma\lambda_n \|\boldsymbol{\theta}_k\|_2 \\ &\quad + (1-\gamma)\lambda_n \|\boldsymbol{\theta}_k\|_1 + \phi_{other}(k), \qquad (2.20)\end{aligned}$$

where $r(-k) = \mathbf{Y} - \sum_{t=1,t\neq k}^{n} \tilde{\mathbf{X}}^{(t)}\boldsymbol{\theta}_t$ and $\phi_{other}(k) = \gamma\lambda_n \sum_{t=1,t\neq k}^{n} \|\boldsymbol{\theta}_i\|_2 + (1-\gamma)\lambda_n \sum_{t=1,t\neq k}^{n} \|\boldsymbol{\theta}_i\|_1$. $\phi_{inner}(\boldsymbol{\theta}_k)$ is a function of coefficients of group k while keeping the coefficients of other groups as constants.

First, we discuss a modified version of SGL algorithm in [30]. We only describe an outline of the algorithm whose details can be found in [30]. For any given λ_n and γ, the algorithm can be described as follows.

Algorithm 3 SGL_RBCD($\mathbf{X}$,$\mathbf{Y}$,p,n)

```
initialize $\boldsymbol{\theta}$ to be a $np \times 1$ zero vector
$r = \mathbf{Y} - \sum_{t=1}^{n} \tilde{\mathbf{X}}^{(t)} \boldsymbol{\theta}_t$
repeat pick $k = 1, \cdots, n$ with probability $\frac{1}{n}$                ▷ outer Loop
    $r(-k) = r + \tilde{\mathbf{X}}^{(k)} \boldsymbol{\theta}_k$
    if the optimal coefficients of group $k$ are identically zeros then
        $\boldsymbol{\theta}_k^{new} = 0$
    else
        initialize $\boldsymbol{\theta}_k^{new}(0)$ to be a $p \times 1$ zero vector
        $i = 0$
        repeat $i = i + 1$                                                 ▷ inner Loop
            $\boldsymbol{\theta}_k^{new}(i) = U(\boldsymbol{\theta}_k^{new}(i-1))$
        until convergence
        $\boldsymbol{\theta}_k^{new} = \boldsymbol{\theta}_k^{new}(i)$
    end if
    $r = r + \tilde{\mathbf{X}}^{(k)}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^{new})$
    $\boldsymbol{\theta}_k^{old} = \boldsymbol{\theta}_k$
    $\boldsymbol{\theta}_k = \boldsymbol{\theta}_k^{new}$
until convergence
     return $\boldsymbol{\theta}$
```

In this algorithm, we have an outer loop which iterates over all groups until convergence, and an inner loop which calculates the optimal coefficients of a particular group while the coefficients of other groups are viewed as constants. The function U in the inner loop is the update function. The function U takes in the old coefficients of group k $\boldsymbol{\theta}_k^{new}(i-1)$ and output the new coefficients of group k $\boldsymbol{\theta}_k^{new}(i)$ until convergence. The function U can be viewed as a black box. For completeness, we expand U as its simplified form in [30]. Let $S(\cdot)$ be the coordinate-wise soft thresholding operator and $(S(z,b))_j := \text{sign}(z_j)(|z_j| - b)_+$, where z is a vector, b is a scalar and $(S(z,b))_j$ is the jth element of the output vector of $S(\cdot)$. Let $l(r_{(-k)}, \boldsymbol{\theta}_k) = \frac{1}{n}\left\|r(-k) - \tilde{\mathbf{X}}^{(k)}\boldsymbol{\theta}_k\right\|_2^2$ be the unpenalized loss function. Then the function U is defined as

$$\begin{aligned} U(\boldsymbol{\theta}_k) &= \left(1 - \frac{t\gamma\lambda_n}{\left\|S(\boldsymbol{\theta}_k - t\nabla l(r_{(-k)}, \boldsymbol{\theta}_k), t(1-\gamma)\lambda_n)\right\|_2}\right)_+ \\ & \quad S(\boldsymbol{\theta}_k - t\nabla l(r_{(-k)}, \boldsymbol{\theta}_k), t(1-\gamma)\lambda_n), \end{aligned}$$

where t is the step size of the gradient method.

The "convergence" is the termination condition, e.g., $\|\boldsymbol{\theta}_k^{new}(i) - \boldsymbol{\theta}_k^{new}(i-1)\|_2 < \epsilon$, $|\phi(\boldsymbol{\theta}_k^{new}) - \phi(\boldsymbol{\theta}_k^{old})| < \epsilon$, etc. In the original algorithm proposed in [30], the outer loop is chosen in a cyclical order. Since the convergence rate of cyclically block-coordinate descent method is unknown except for some special cases [73], in the algorithm described above, we modify the outer loop of algorithm in [30] to the randomized block-coordinate descent (RBCD) method in [74]. We will use Algorithm 2.4 in our simulation.

Although Algorithm 2.4 is easy to implement, it is difficult to analyze its complexity due to its inexact nature of the inner loop. In the following, we introduce the modified Signed Single Line Search (SSLS) algorithm in [31], which is more amenable to complexity analysis.

Algorithm 4 SSLS_RBCD($\mathbf{X}$,$\mathbf{Y}$,p,n)

```
initialize $\boldsymbol{\theta}$ to be a $np \times 1$ zero vector
$r = \mathbf{Y} - \sum_{t=1}^{n} \tilde{\mathbf{X}}^{(t)}\boldsymbol{\theta}_t$
repeat pick $k = 1, \cdots, n$ with probability $\frac{1}{n}$          ▷ outer Loop
    $r(-k) = r + \tilde{\mathbf{X}}^{(k)}\boldsymbol{\theta}_k$
    $\boldsymbol{\theta}_k^{new} = \text{argmin}_{\boldsymbol{\theta}_k} \{\phi_{inner}(\boldsymbol{\theta}_k)\}$          ▷ inner Loop
    $r = r + \tilde{\mathbf{X}}^{(i)}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^{new})$
    $\boldsymbol{\theta}_k = \boldsymbol{\theta}_k^{new}$
until convergence
    return $\boldsymbol{\theta}$
```

The original SSLS algorithm has a similar structure as the SGL algorithm in [30]. In the description above, we make the same modification, i.e., change the cyclical order to random order in the outer loop. The main difference between Algorithm 2.4 and 2.4 lies in the inner loop. In the inner loop, SSLS explicitly solves the optimal group coefficients for one group while keeping the coefficients of other groups as constants. The main result is in the following Proposition.

Proposition 6. *For an error tolerance ϵ and a given constant confident level $\rho \in (0, 1)$, the complexity of the randomized block-coordinate descent method version of SSLS for the worst case is $O(n^2\lambda_n^{-2})$ that guarantees*

$$P(\phi(\hat{\boldsymbol{\theta}}) - \phi^* \leq \epsilon) \geq 1 - \rho. \tag{2.21}$$

Proof. Please see Appendix A.7. □

The complexity relies on λ_n. For example, if $\lambda_n = \frac{1}{\sqrt{\ln n}}$ and $\delta = \frac{1}{\sqrt[3]{\ln n}}$, a valid choice for Proposition 1-4 under A1-A4, then the complexity is $O(n^2 \ln n)$, which is better than the complexity of DP approach $O(n^3)$.

2.5 Numerical Results

Our simulation is based on the Algorithm 1, a slightly modified version of the algorithm in [30] that has an R implementation in the package SGL. We first test our algorithm on synthesised data. We select $p = 20$, and set the number of nonzero coefficients to be 4. In particular, we set the first four coefficients to be nonzero in each $\boldsymbol{\beta}_t$. In our simulation, we set $\gamma = 0.927$, $n = 1000$, $K^* = 3$, and the real change-points are at $100, 300, 900$. The first four coefficients of $\boldsymbol{\beta}_t$s are all 2 for points $t = 1, \cdots, 99$ and for $t = 100, \cdots, 299$, and are all -2 for $t = 300, \cdots, 899$ and for $t = 900, \cdots, 1000$. Each $x_{t,m} \sim \mathcal{N}(0, 4)$, and the noise $\epsilon_t \sim \mathcal{N}(0, 0.01)$.

In our simulation result figures, x-axis represents the locations from 1 to n, and y-axis represents whether the data point at each location is an estimate change-point (1 means that it is an estimated change-point, i.e., $\hat{\boldsymbol{\theta}}_t \neq \mathbf{0}$, while 0 means $\hat{\boldsymbol{\theta}}_t = \mathbf{0}$.).

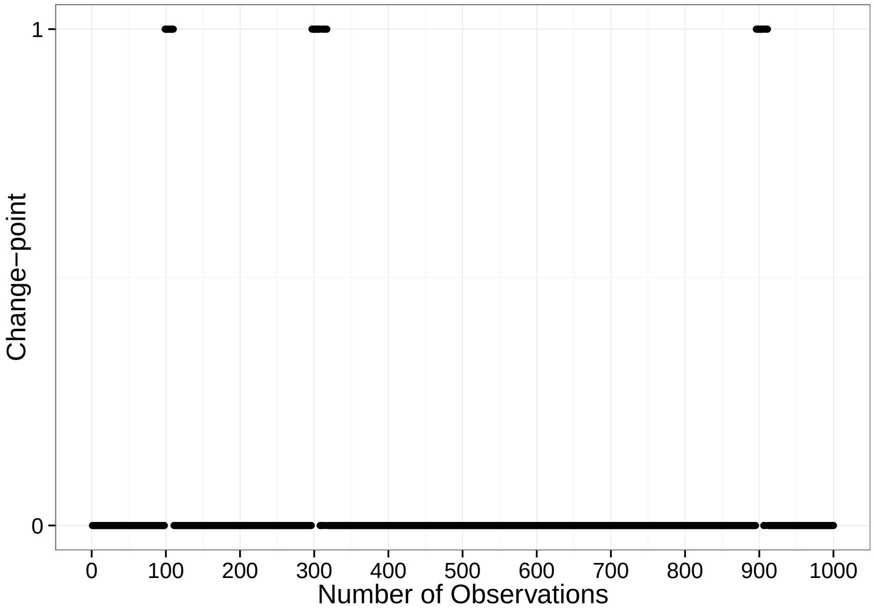

Figure 2.2: Change-points locations estimation using SGL, $\lambda = 0.003778942$.

From Figure 2.2, we can see that most of the estimated $\hat{\boldsymbol{\theta}}_t$ are zero vectors. Furthermore, the nonzero vectors are clustered around the true change-points. This implies that our approach can successfully identify the locations of change-points.

We also examine the regularization path of $\hat{\boldsymbol{\theta}}_{300}$ by choosing different values of λ_n. If we select λ_n large enough, then the penalty term will dominate the SGL and encourages the sparsity, thus the coefficients are tending to be all zeros. If we decrease λ_n, the least square term becomes more and more dominant, thus the sparsity of coefficients will decrease. On the other hand, from our asymptotic results, we know that the accuracy will increase. Hence, in practice, we need to set λ_n properly to balance between accuracy and sparsity. From regularization path, we find that once λ_n is properly chosen, our algorithm can properly identify the important coefficients.

Here we compare our approach with the DP based approach. Figure 2.3 illustrates the estimated change point using the DP based approach when K_{max} is set to be 3, the true value of K^*. It shows that the change-points location estimates are accurate if we know K^*. However, as discussed in Section 1.1.2, if K^* is unknown and only $K_{\max}$ is known, the DP based approach will return K_{max} change-points. Figure 2.4 shows the change-points estimates using the DP based approach when K_{max} is set to be 20. From the figure, we can see that the returned change-points estimates do not concentrate around the true change points and hence do not provide accurate estimates of the true underlying change-points. Furthermore, the coefficients of all the results by DP approach do no possess sparse structure, which means that most or even all of the coefficients are nonzero while the results of our SGL based approach possess sparsity.

Next, we test our approach on real weather data collected by NOAA. We use NCEP/NCAR Reanalysis 1 Surface Monthly Mean dataset [75]. The dataset records monthly means of precipitation for 1948-present for all locations on the globe, and each locations has $2.5^\circ \times 2.5^\circ$ resolution. Our goal is to find change-points in climate models for different locations. We pick 5 target locations, i.e., Eastern USA, Brazil, Western USA,

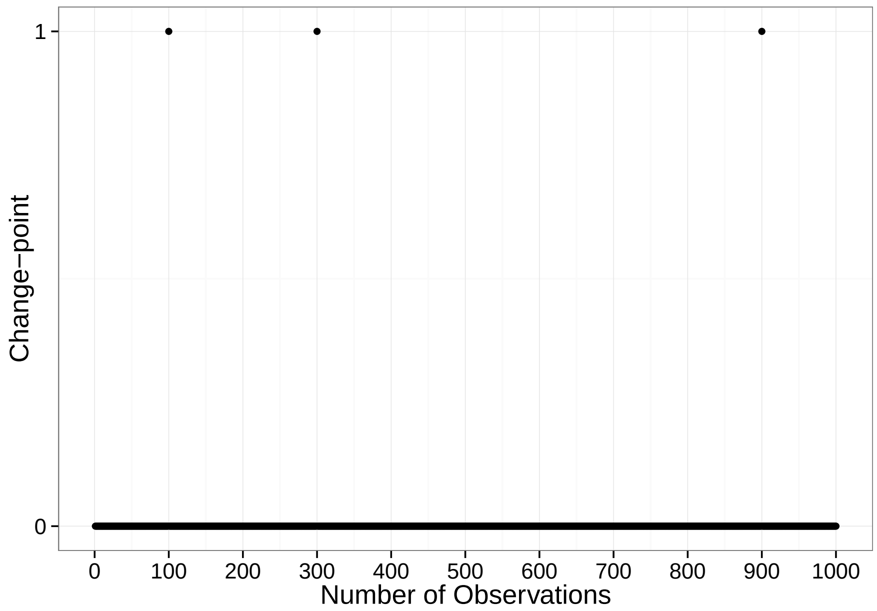

Figure 2.3: Change-points locations estimation using DP, $K_{max} = 3$.

South Africa and India because of their diverse geological properties. The parameters of these target locations are considered as $\mathbf{Y}$ in our model. Then we pick 40 locations near the target locations as the $\{\mathbf{x}_t\}$ in our model. For each location, we pick the first 400 data. Then we concatenate the data $\mathbf{Y}$ and $\{\mathbf{x}_t\}$ for different locations. Hence, we have $n = 5 \times 400 = 2000$ and $p = 40$. And in our concatenated data, the first segment, i.e., $1 \leq n \leq 400$, the model describes the relationship between Eastern USA and our 40 data locations, the second segment, i.e., $401 \leq n \leq 800$, the model describes the relationship between Brazil and our 40 data locations, and so on. We choose the precipitation as the parameter to be investigated in the model, and $\gamma = 0.8634729$ in our simulation.

Figure 2.5 shows the l_1 norm of our result. From Figure 2.5 we can see the inter-group sparsity of the result. Furthermore, the estimated change-points are clustered around the true change-points.We also examine the regularization path. For the first estimated interval, when $\lambda = 0.01267545$, coefficients at indices $9 - 16$ and $25 - 32$ are zero, that is

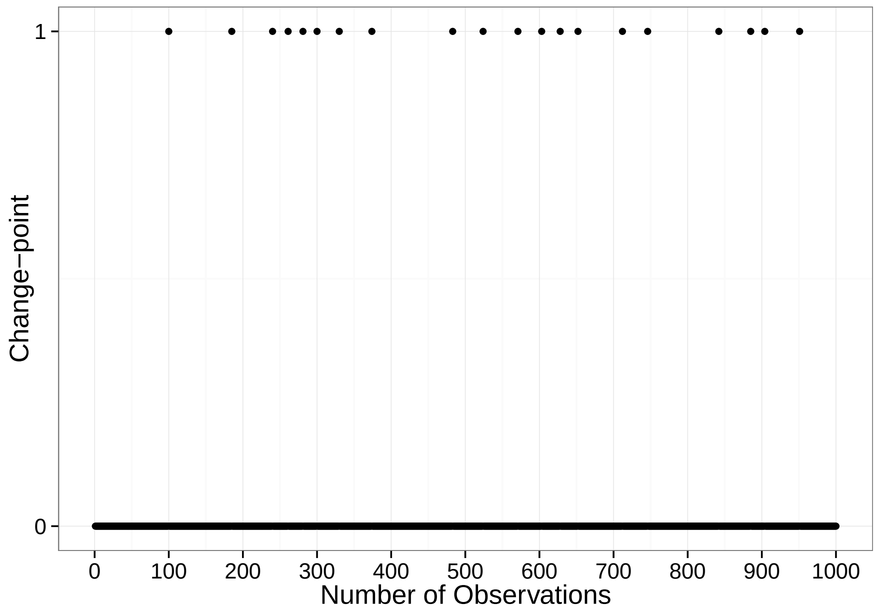

Figure 2.4: Change-points locations estimation using DP, $K_{max} = 20$.

16 coefficients out of 40 are zero, which show the sparsity within the group. Furthermore, these coefficients corresponds to locations near eastern US, western US and India. Since the data in Y of segment $n = 1$ to $n = 400$ is from eastern US, the result above indicates that precipitation of eastern US has a higher correlation with precipitation of locations near eastern US, western US and India than precipitation of Brazil and South Africa, since eastern US, western US and India are all located in northern hemisphere and near heavily rained regions which is consistent of [76].

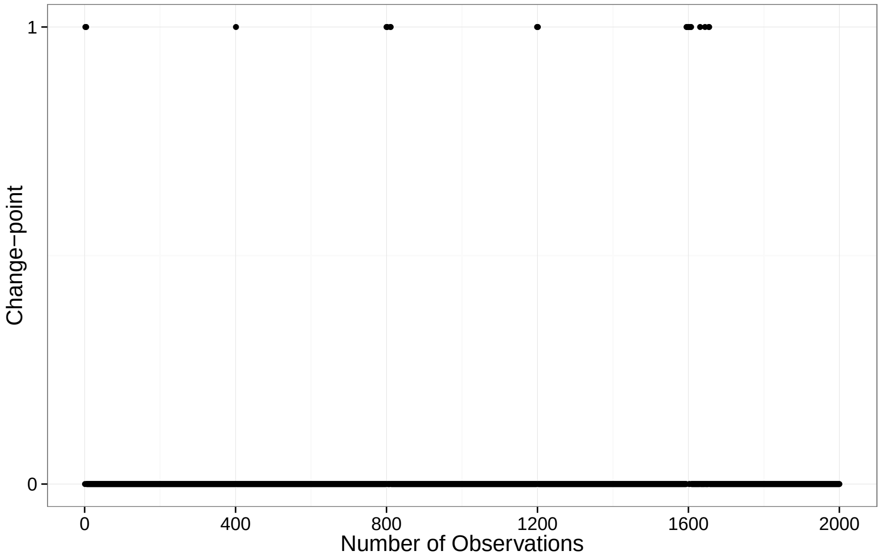

Figure 2.5: Change-points locations estimation using SGL, $\lambda = 0.01546122$.

Chapter 3

High Dimensional Change-points Inference

In this chapter, we extend our analysis in low dimensional linear regression models to high dimensional setting and further extend it to GLM. In Section 3.1, we describe the model under high dimensional setting. In Section 3.2, we prove the consistency and properties of the solution of our approach. In Section 3.3, we extend our study to generalized linear models. In Section 3.4, we provide numerical examples to illustrate the performance of our approach.

3.1 Model

3.1.1 Problem Formulation

Here we consider the linear regression model in (2.1). Since this section focuses on high dimensional case, here $\boldsymbol{\beta}_t^* \in \mathbb{R}^p$ is a sparse coefficients vector with sparse level s and p/n does not go to zero as $n \to \infty$.

Here we perform the same transformation as in Section 2.2. We note that $\boldsymbol{\theta}^*$ has a very unique sparsity structure. In particular, as there are only K^* change-points among n observations, we can see that there are only $K^* + 1$ sub-vectors of $\boldsymbol{\theta}^*$ are nonzero. We call this group-wise sparsity. Furthermore, even for those non-zero sub-vectors $\boldsymbol{\theta}_t^*$, as both $\boldsymbol{\alpha}_{k+1}^*$ and $\boldsymbol{\alpha}_k^*$ are s-sparse vectors, most of the entries of the non-zero subvector $\boldsymbol{\theta}_t^*$ are zero. We call this within group sparsity. Hence, $\boldsymbol{\theta}^*$ possesses both group-wise and within group sparsity structure.

Motivated by this unique sparsity structure of $\boldsymbol{\theta}^*$, we propose to solve the multiple change-points estimation problem via solving the same optimization problem as (2.7)

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{np}} \varphi(\boldsymbol{\theta}) := \underbrace{\frac{1}{n}\|\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\theta}\|_2^2}_{\mathcal{L}(\boldsymbol{\theta})} + \lambda_n \underbrace{\left[\gamma \sum_{t=1}^{n} \|\boldsymbol{\theta}_t\|_2 + (1-\gamma)\|\boldsymbol{\theta}\|_1\right]}_{\mathcal{R}(\boldsymbol{\theta})}, \tag{3.1}$$

in which $\mathcal{L}(\boldsymbol{\theta})$ is referred as the loss function, the function $\mathcal{R}(\boldsymbol{\theta})$ is referred as the regularization penalty function, λ_n is the regularization penalty weight, and $\gamma \in (0, 1)$ adjusts the relative weight between the two terms in $\mathcal{R}(\boldsymbol{\theta})$. Here, the design of the penalty term $\mathcal{R}(\boldsymbol{\theta})$ is motivated the unique sparsity structure of $\boldsymbol{\theta}^*$ discussed above. In particular, the term $\sum_{t=1}^n \|\boldsymbol{\theta}_t\|_2$ is used to encourage the group-wise sparsity in the solution, and the term $\|\boldsymbol{\theta}\|_1$ is used to encourage within group sparsity in the solution.

As the analysis in Chapter 2 relies crucially on p being fixed, the analysis approach used in Chapter 2 does not apply in this high-dimension case. In this thesis, we will provide new analysis that works for the high-dimension case.

As there is a one-to-one correspondence between $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, the optimization problem in (3.1) can be equivalently written as

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{np}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_n \left[\gamma \sum_{t=1}^{n} \|\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}\|_2 + (1-\gamma) \sum_{t=1}^{n} \|\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}\|_1\right]. \tag{3.2}$$

Let $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ denote the optimal solution of (3.1) and (3.2) respectively. $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ can be transformed to each other by noticing

$$\begin{aligned}\hat{\boldsymbol{\beta}}_t &= \sum_{i=1}^{t} \hat{\boldsymbol{\theta}}_i, && (3.3)\\ \hat{\boldsymbol{\theta}}_t &= \hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_{t-1}. && (3.4)\end{aligned}$$

Furthermore, the locations of estimated change-points are those t's such that $\hat{\boldsymbol{\theta}}_t$ is nonzero, and the estimated total number of change-points $\hat{K}$ is the total number of nonzero $\hat{\boldsymbol{\theta}}_t$'s.

Let

$$\begin{aligned}\boldsymbol{\Delta}_{\boldsymbol{\theta}} &:= \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*,\\ \boldsymbol{\Delta}_{\boldsymbol{\beta}} &:= \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*,\end{aligned}$$

be the estimation error vectors. Directly from the definition above, we have

$$\mathbf{X}\boldsymbol{\Delta}_{\boldsymbol{\beta}} = \tilde{\mathbf{X}}\boldsymbol{\Delta}_{\boldsymbol{\theta}}. \qquad (3.5)$$

3.1.2 Assumptions on Data

Throughout the chapter, we make following assumptions:

A1. $\mathbf{x}_t$ is generated from the distribution $\mathcal{N}(0, \boldsymbol{\Sigma})$, and $0 < l \leq \mu_{\min}(\boldsymbol{\Sigma}) \leq \mu_{\max}(\boldsymbol{\Sigma}) \leq L < \infty$, where l and L are constants independent of n and p.

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

A1 indicates that our design matrix is Gaussian ensemble as discussed in [8, 47, 48]. A2 puts a constraint on the covariance matrix of the Gaussian ensemble. Notice that the diagonal elements of $\boldsymbol{\Sigma}$ are variances of elements in $\mathbf{x}_t$. A2 means that the variance of each element of $\mathbf{x}_t$ is upper bounded by a constant.

In the next section, we will first derive general results without making particular assumptions on $I_{\min}$ and $J_{\min}$ which are defined in (2.8) and (2.9). We will then simplify the results when particular assumptions on these quantities are made. Those particular assumptions will be introduced in Section 3.2.3.

3.2 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.

3.2.1 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 Ξ be a norm on $\mathbb{R}^m$, its dual norm Ξ^* is given by*

$$\Xi^*(\boldsymbol{\theta}) = \sup_{\boldsymbol{v}\in\mathbb{R}^m\backslash\{\mathbf{0}\}} \frac{< \boldsymbol{v}, \boldsymbol{\theta} >}{\Xi(\boldsymbol{v})} = \sup_{\Xi(\boldsymbol{v})\leq 1} < \boldsymbol{v}, \boldsymbol{\theta} >,$$

in which $< \cdot, \cdot >$ denotes the inner product of two vectors.

Lemma 1. *Let $\hat{\boldsymbol{\theta}}$ be an optimal solution to the optimization problem* (3.1) *when we choose*

$$\lambda_n \geq 2\mathcal{R}^* \left(\nabla\mathcal{L}(\boldsymbol{\theta}^*)\right),$$

in which $\mathcal{R}^*$ is the dual norm of $\mathcal{R}$, then we have

$$\mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) \leq 4\mathcal{R}(\boldsymbol{\theta}^*). \tag{3.6}$$

Proof. Please see Appendix A.8.1. □

Lemma 2. *Let $\mathbf{U} \in \mathbb{R}^{m_1 \times m_2}$, suppose each row $\mathbf{U}_{i,\cdot}, i = 1, \cdots, m_1$ are independently generated using $\mathcal{N}(0, \boldsymbol{\Sigma}_{\mathbf{U}})$ with $0 < l_{\mathbf{U}} \leq \mu_{\min}(\boldsymbol{\Sigma}_{\mathbf{U}}) \leq \mu_{\max}(\boldsymbol{\Sigma}_{\mathbf{U}}) \leq L_{\mathbf{U}} < \infty$, then we have*

$$\sqrt{l}\|\boldsymbol{v}\|_2 \;\; \leq \;\; \|\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{v}\|_2 \leq \sqrt{L}\|\boldsymbol{v}\|_2. \tag{3.7}$$

Furthermore, with a probability at least $1 - c^{'} \exp(-cm_1)$, in which c, $c^{'}$ are universal positive constants from [47, Theorem 1], we have

$$\frac{1}{m_1}\|\mathbf{U}\boldsymbol{v}\|_2^2 \;\; \geq \;\; \frac{l}{32}\|\boldsymbol{v}\|_2^2 - \frac{9}{4}\sqrt{L}\rho(\boldsymbol{\Sigma}_{\mathbf{U}})\sqrt{\frac{\log m_2}{m_1}}\|\boldsymbol{v}\|_2\|\boldsymbol{v}\|_1, \tag{3.8}$$

hold for all $\boldsymbol{v} \in \mathbb{R}^{m_2}$ simultaneously.

Proof. Please see Appendix A.8.2. □

3.2.2 Results for General Models

In this subsection, we present our main results regarding the estimation error for the general model specified in Section 3.1. These results will be further simplified in Section 3.2.3 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 $\hat{\boldsymbol{\theta}}$ be the optimal value obtained from the optimization problem (3.1), then we have the following inequality

$$\varphi(\hat{\boldsymbol{\theta}}) \leq \varphi(\boldsymbol{\theta}^*). \tag{3.9}$$

The basic idea of the proofs is to show that, for (3.9) to hold, the conclusions in the following propositions must hold.

Before presenting the results, we define a quantity that will be frequently used

$$\delta_n := \frac{64}{lJ_{\min}^2}\left(\frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) + 2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*)\right). \tag{3.10}$$

In all results presented in this subsection, we assume that A1-A2 hold and we choose

$$2\mathcal{R}^*\left(\nabla\mathcal{L}(\boldsymbol{\theta}^*)\right) \leq \lambda_n < \frac{1}{2\mathcal{R}(\boldsymbol{\theta}^*)}\left(\frac{I_{\min}}{2n}\frac{lJ_{\min}^2}{64} - \frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*)\right) \tag{3.11}$$

Before moving further, we would like to comment on (3.11), which puts lower and upper bounds on λ_n. It is reasonable to have lower and upper bounds on λ_n. If λ_n is too small, the solution of problem (3.1) 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 λ_n is too large, the solution of problem (3.1) will not be able to capture change-points, as the estimated vector will become sparser as λ_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 frequently arises in the detailed analysis. The particular form of upper bound comes from the application of (3.8) in the detailed analysis. It is natural to ask whether such λ_n exists and if so how to choose such λ_n without knowing the values of quantities such as $\boldsymbol{\theta}^*$ and $I_{\min}$ etc. We will answer these questions in Section 3.2.3 which shows such λ_n exists under some mild assumptions and shows how to choose one.

Here we make an additional mild assumption on $I_{\min}$ and δ_n.

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 3.2.3. Furthermore, noticing that since we choose λ_n satisfying (3.11), 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.

Proposition 7. *If $\hat{K} = K^*$, then we have*

$$\frac{\max_{1 \le k \le K^*} |\hat{t}_k - t_k^*|}{n} \le \delta_n. \tag{3.12}$$

Proof. Please see Appendix A.9.1 for details. □

Proposition 7 shows that, if $\hat{K} = K^*$ and δ_n goes to zero as n and p increase, then the relative errors of estimated change-points locations diminish.

Proposition 8. *If $\hat{K} = K^*$, then for $k \in [K^* + 1]$,*

$$\|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\|_2^2 \le \frac{32n}{(I_{\min} - 2n\delta_n)l}\left(\frac{36}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) + 2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*)\right) \tag{3.13}$$

Proof. Please see Appendix A.9.2. □

The condition for Proposition 8 is the same as the conditions for Proposition 7. As will be shown in Section 3.2.3, if more details of the model is provided, the result can be further simplified.

Proposition 9.

$$\hat{K} \ge K^*. \tag{3.14}$$

Proof. Please see Appendix A.9.3. □

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

The results in Proposition 7 and 8 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. The following proposition is parallel to Proposition 7 for the case $\hat{K} > K^*$.

Proposition 10.

$$\frac{\varepsilon\left(\hat{\mathbb{T}}_{\hat{K}}||\mathbb{T}^*\right)}{n} \leq \delta_n.$$

Proof. Please see Appendix A.9.4. □

Proposition 10 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 squeal. 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

$$\begin{aligned}\mathbb{T}_{isolate} \quad := \quad & \left\{\exists \hat{t}_j \in [t_k^*, t_{k+1}^* - 1] \text{ for some } k \in [K^*+1] \text{ s.t. } \min\left\{\hat{t}_{j+1}, t_{k+1}^*\right\} - \hat{t}_j \geq n\delta_n,\right. \\ & \left.\hat{t}_j - \max\left\{\hat{t}_{j-1}, t_k^*\right\} \geq n\delta_n, \|\hat{\boldsymbol{\alpha}}_{j+1} - \hat{\boldsymbol{\alpha}}_j\|_2 > J_{\min}\right\},\end{aligned}$$

where δ_n is the nonnegative quantity defined in (3.10). The estimated change-points described in $\mathbb{T}_{isolate}$ are the isolated change-points with a big change in coefficients. The variable δ_n is used to measure the distance between the isolated change-point and its nearest true or estimated change-points. Figure 3.1 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.

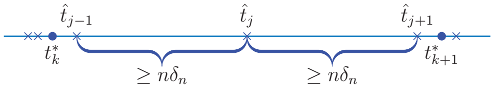

Figure 3.1: Illustration of an isolated change-point $\hat{t}_j$ with $\hat{t}_{j-1} \geq t_k^*$, $\hat{t}_{j+1} \leq t_{k+1}^*$, $\hat{t}_{j+1} - \hat{t}_j \geq n\delta_n$, $\hat{t}_j - \hat{t}_{j-1} \geq n\delta_n$, and $\|\hat{\boldsymbol{\alpha}}_{j+1} - \hat{\boldsymbol{\alpha}}_j\|_2 > J_{\min}$.

Proposition 11. $\mathbb{T}_{isolate}$ *does not happen.*

Proof. Please see Appendix A.9.5. □

Proposition 11 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 very good estimate of the linear coefficients*.

To make the statements above precise, we first introduce some definitions. Let $\xi_k^* := [t_{k-1}^*, t_k^*)$ be the kth true interval. Let $\hat{\xi}_i := [\hat{t}_{i-1}, \hat{t}_i)$ be the ith estimated interval, and $\hat{I}_i$ be the length of the ith estimate interval. We define the set

$$\Upsilon(\varsigma_n) := \{i \in [\hat{K}+1] | \hat{\xi}_i \subseteq \xi_k^* \text{ for some } k \in [K^*+1] \text{ and } \hat{I}_i > n\varsigma_n\},$$

where ς_n is some nonnegative quantity. In the following, we take $n\varsigma_n = \frac{I_{\min}}{10}$. Then $\Upsilon(\frac{I_{\min}}{10n})$ contains all i's such that the ith interval between estimated change-points is longer than $\frac{I_{\min}}{10}$. Here we pick $n\varsigma_n = \frac{I_{\min}}{10}$ since we focus on long estimated intervals that are not too short compared to $I_{\min}$. Actually, for our result to hold, $\frac{I_{\min}}{c}$ would suffice, where $c > 0$. Let $\boldsymbol{\alpha}^*_{k(i)}$ be the coefficients of "long" intervals containing $\hat{\xi}_i$ for $i \in \Upsilon(\varsigma_n)$.

Proposition 12 (Bounds for long intervals)**.** *For $i \in \Upsilon\left(\frac{I_{\min}}{10n}\right)$, we have*

$$\|\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}^*_{k(i)}\|_2^2 \leq \frac{320n}{I_{\min}l}\left(\frac{36}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) + 2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*)\right). \quad (3.15)$$

Proof. Please see Appendix A.9.6. □

3.2.3 Simplified Results with Knowledge of Model Details

The results provided in Section 3.2.2 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. $I_{\min} = \Theta(n)$; $J_{\min} = \Omega(1)$; $\mathcal{R}(\boldsymbol{\theta}^*) = O(s)$; $s^2\sqrt{\frac{\log p}{n}} \to 0$ as $n \to \infty$.

Since $I_{\min} \leq n$, then $I_{\min} = \Theta(n)$ means that intervals between true change-points are comparable with the total number of samples, i.e., the intervals between change-points are not too small. $J_{\min} = \Omega(1)$ means the growth order of $J_{\min}$ 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 $(K^* + 1)s$ nonzero elements in $\boldsymbol{\theta}^*$ and K^* is a constant, we can assume $\mathcal{R}(\boldsymbol{\theta}^*) = O(s)$ if we know each entry of the coefficients are bounded. $s^2\sqrt{\frac{\log p}{n}} \to 0$ 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 λ_n has very desirable properties that will enable us to simply the results obtained in Section 3.2.2.

Lemma 3. *If A1-A3 holds and we choose* $\lambda_n = \frac{8\sigma\sqrt{5\rho_c}}{1-\gamma}\sqrt{\frac{\log np}{n}}$*, 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* (3.13) *can be upper bounded by*
$$c_8 s^2 \sqrt{\frac{\log p}{n}}.$$
- *Condition* (3.11) *is satisfied with probability at least* $1 - \frac{2}{np} - np\exp(-n)$.

Proof. Please see Appendix A.10. □

Using Lemma 3, it is easy to see that assumption B1 is satisfied for large n and we can rewrite Proposition 7-12 into simplified versions in the following remarks. In the following remarks, we assume A1-A3 hold, n is large enough and we choose $\lambda_n = \frac{8\sigma\sqrt{5\rho_c}}{1-\gamma}\sqrt{\frac{\log np}{n}}$.

Remark 7. *If* $\hat{K} = K^*$*, then*

$$\frac{\max_{1\leq k\leq K^*} |\hat{t}_k - t_k^*|}{n} \leq c_4 s^2 \sqrt{\frac{\log p}{n}},$$

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

Note that a probability $1 - \frac{2}{np} - np\exp(-n)$ is introduced compared to Proposition 7. This is due to the choice of $\lambda_n = \frac{8\sigma\sqrt{5\rho_c}}{1-\gamma}\sqrt{\frac{\log np}{n}}$, which satisfies condition (3.11) with a probability as shown in Lemma 3. Since we require (3.11) to be true for Propositions 7-12 to hold, hence in the following remarks, we add an extra probability term to each proposition.

Remark 8. *If $\hat{K} = K^*$, then*

$$\sum_{k=1}^{K^*+1} \|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\|_2^2 \leq c_8 s^2 \sqrt{\frac{\log p}{n}},$$

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

Remark 9.

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

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

Remark 10.

$$\frac{\epsilon\left(\hat{\mathbb{T}}_{\hat{K}} || \mathbb{T}^*\right)}{n} \leq c_4 s^2 \sqrt{\frac{\log p}{n}},$$

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

Remark 11. $\mathbb{T}_{isolate}$ *does not happen with probability at least $1 - \frac{2}{np} - np \exp{(-n)}$.*

Remark 12. *For $i \in \Upsilon\left(\frac{I_{\min}}{10n}\right)$,*

$$\|\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_{k(i)}^*\|_2^2 \leq c_{10} s^2 \sqrt{\frac{\log p}{n}},$$

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

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

3.3 Generalized Linear Models

In this section, we extend our results obtained in linear models to generalized linear models (GLM) [8]. In GLM, y_t and $\mathbf{x}_t$ are related through the linear coefficient $\boldsymbol{\beta}_t^*$ as

$$P(y_t|\mathbf{x}_t, \boldsymbol{\beta}_t^*) \propto \exp\left\{\frac{y_t\langle\boldsymbol{\beta}_t^*, \mathbf{x}_t\rangle - \Phi\left(\langle\boldsymbol{\beta}_t^*, \mathbf{x}_t\rangle\right)}{c}\right\}.$$

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 [8], it is assumed that there is no change in the underlying model, and hence $\boldsymbol{\beta}_t^*$ is assumed to be a constant for all t. We will focus on GLM with change-points, i.e., we assume $\boldsymbol{\beta}_t^*$ is not constant anymore. In particular, similar to Section 3.1, we assume that the parameter $\boldsymbol{\beta}_t^*$ is piece-wise constant with respect to t. More specifically, let K^* denote the number of changes in $\boldsymbol{\beta}_t^*$, then for $k \in [K^*]$,

$$\boldsymbol{\beta}_t^* = \boldsymbol{\alpha}_k^*, \quad \text{for } t \in [t_{k-1}^*, t_k^*), \text{ and } \boldsymbol{\alpha}_k^* \neq \boldsymbol{\alpha}_{k-1}^*,$$

where $t_0^* := 1$ and $t_{K^*+1}^* := n + 1$ by convention. Following the same approach as in the previous sections, we let $\boldsymbol{\theta}_t := \boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}$, and let $\boldsymbol{\theta}^* \in \mathbb{R}^{np}$ denote the vector formed by concatenating each $\boldsymbol{\theta}_t^*$, $t = 1, \cdots, n$.

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

Using the same linear transformation and follow the same reasoning in Section 2.1, we propose to perform the change-points estimation problem by solving

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^{np}}{\text{argmin}} \left\{ \underbrace{\frac{1}{n}\sum_{t=1}^{n}\left(-y_t\langle\boldsymbol{\theta},\tilde{\mathbf{x}}_t\rangle + \Phi\left(\langle\boldsymbol{\theta},\tilde{\mathbf{x}}_t\rangle\right)\right)}_{\mathcal{L}(\boldsymbol{\theta})} + \lambda_n \underbrace{\left[\gamma\sum_{t=1}^{n}\|\boldsymbol{\theta}_t\|_2 + (1-\gamma)\|\boldsymbol{\theta}\|_1\right]}_{\mathcal{R}(\boldsymbol{\theta})} \right\}, \tag{3.16}$$

in which $\tilde{\mathbf{x}}_t$ is the tth row of $\tilde{\mathbf{X}}$ as defined in (2.4).

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

Similar to the previous section, we define

$$\delta_n := \frac{64}{\mu l J_{\min}^2}\left(\mu\frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) + 6\lambda_n\mathcal{R}(\boldsymbol{\theta}^*)\right). \tag{3.17}$$

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

$$2\mathcal{R}^*\left(\nabla\mathcal{L}(\boldsymbol{\theta}^*)\right) \leq \lambda_n \leq \frac{1}{6\mathcal{R}(\boldsymbol{\theta}^*)}\left(\frac{I_{\min}}{2n}\frac{\mu l J_{\min}^2}{64} - \mu\frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*)\right), \tag{3.18}$$

which implies that $\delta_n < \frac{I_{\min}}{2n}$. Under these assumptions, the following propositions show that the results obtained for the linear regression case also hold for GLM.

Proposition 13. *If $\hat{K} = K^*$, then we have*

$$\frac{\max_{1\leq k\leq K^*}|\hat{t}_k - t_k^*|}{n} \leq \delta_n. \tag{3.19}$$

Proposition 14. *If $\hat{K} = K^*$, then for $k \in [K^* + 1]$,*

$$\|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\|_2^2 \leq \frac{32n}{\mu l \left(I_{\min} - 2n\delta_n\right)} \left(\mu \frac{72}{(1-\gamma)^2} \sqrt{L} \rho(\boldsymbol{\Sigma}) \sqrt{\frac{\log p}{n}} \mathcal{R}^2(\boldsymbol{\theta}^*) + 6\lambda_n \mathcal{R}(\boldsymbol{\theta}^*)\right). \tag{3.20}$$

Proposition 15.

$$\hat{K} \geq K^*. \tag{3.21}$$

Proposition 16.

$$\frac{\epsilon\left(\hat{\mathbb{T}}_{\hat{K}} || \mathbb{T}^*\right)}{n} \leq \delta_n. \tag{3.22}$$

Proposition 17. $\mathbb{T}_{isolate}$ *does not occur.*

Proposition 18 (Bounds for long intervals)**.** *For $i \in \Upsilon(I_{\min}/10)$*

$$\|\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_{k(i)}^*\|_2^2 \leq \frac{320n}{\mu l I_{\min}} \left(\mu \frac{72}{(1-\gamma)^2} \sqrt{L} \rho(\boldsymbol{\Sigma}) \sqrt{\frac{\log p}{n}} \mathcal{R}^2(\boldsymbol{\theta}^*) + 6\lambda_n \mathcal{R}(\boldsymbol{\theta}^*)\right). \tag{3.23}$$

3.4 Numerical Simulation

Our simulation is based on Algorithm in Chapter 2, which has an R implementation in the package ChangePointCalc [77].

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 $\boldsymbol{\beta}_t$. In our simulation, we set $\gamma = 0.927$, $n = 800$, $K^* = 3$, and the real change-points are at $101, 301, 701$. For each $\boldsymbol{\alpha}_k^*, k = 1, \cdots, 4$, the first s coefficients are drawn independently from uniform distribution $\mathcal{U}[-2, 2]$ 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 3.2, 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 that it is an estimated change-point, i.e., $\hat{\boldsymbol{\theta}}_t \neq \mathbf{0}$, while 0 means $\hat{\boldsymbol{\theta}}_t = \mathbf{0}$.).

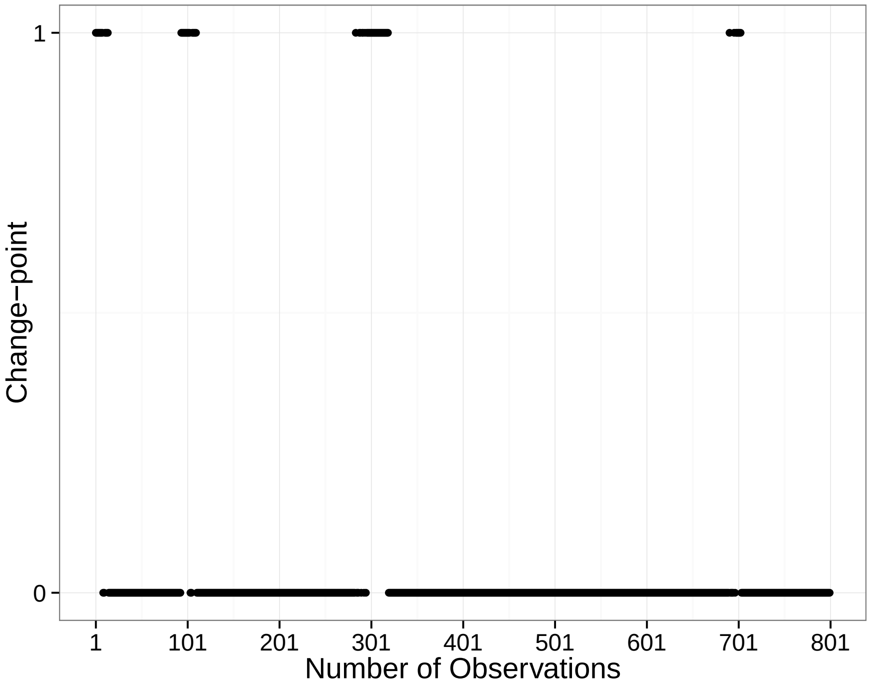

Figure 3.2: Change-points locations estimation using SGL for ordinary linear regression, $\lambda_n = 0.009125759$.

In Figure 3.3, y-axis represents the l_2-norm of each $\hat{\boldsymbol{\theta}}_t$, $t \in [n]$.

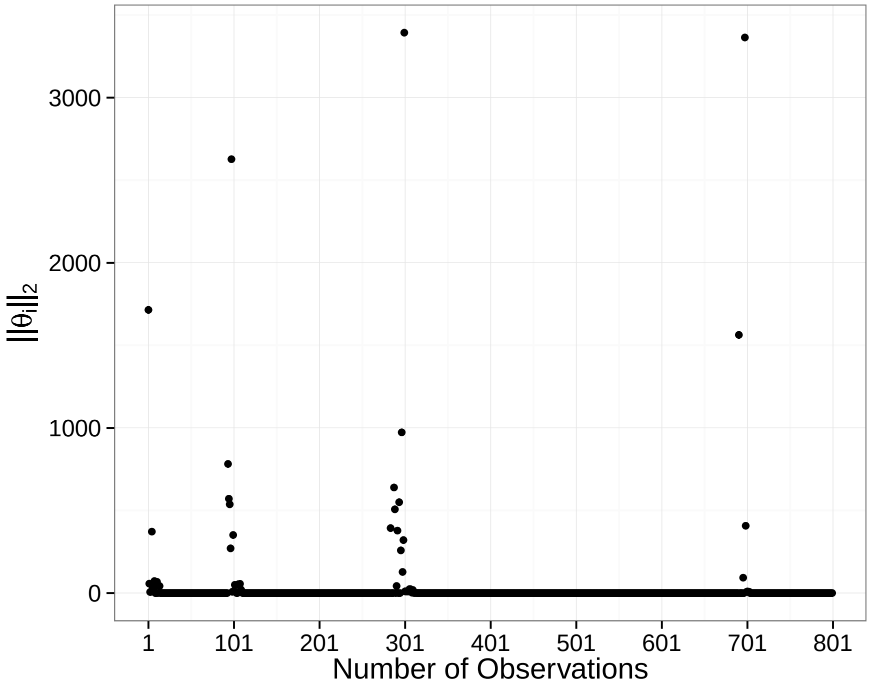

Figure 3.3: l_2-norm of $\hat{\boldsymbol{\theta}}_t$, $t \in [n]$ for ordinary linear regression, $\lambda_n = 0.009125759$.

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 theoretic results obtained.

Next, we show our simulation results for logistics 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 3.4 and 3.5, we again can 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.

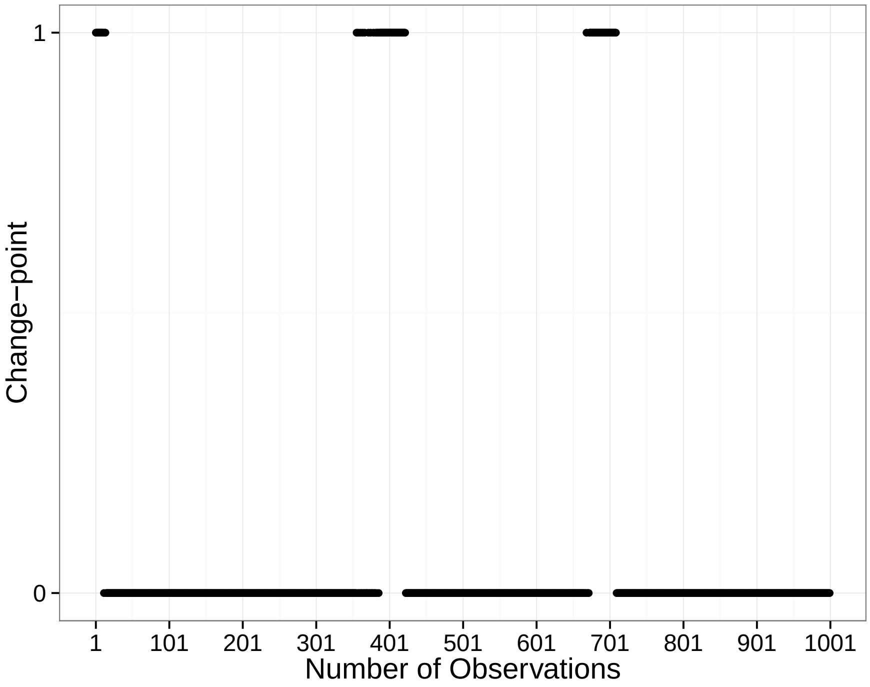

Figure 3.4: Change-points locations estimation using SGL for logistic regression, $\lambda_n = 0.002369588$.

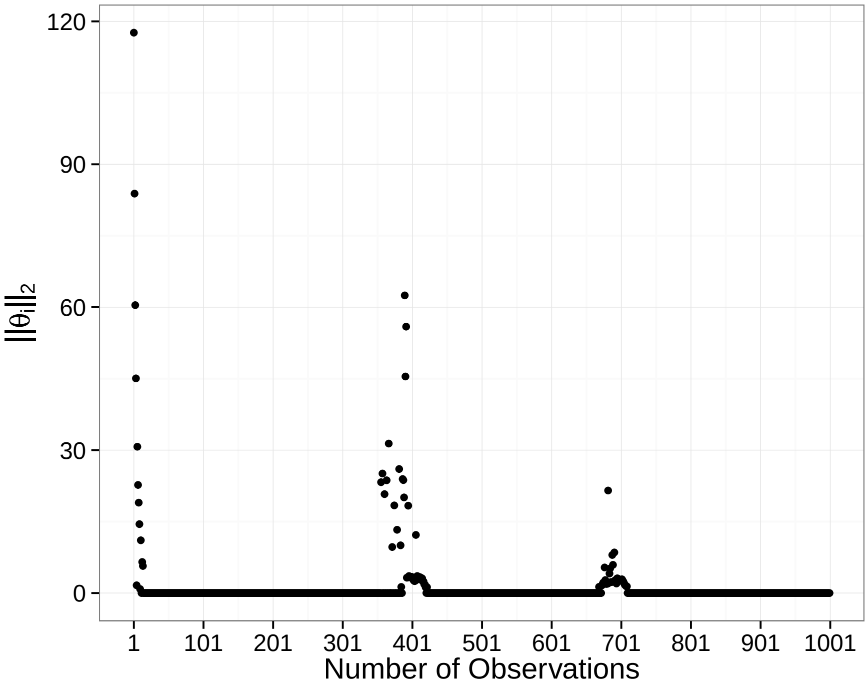

Figure 3.5: l_2-norm of $\hat{\boldsymbol{\theta}}_t$, $t \in [n]$ for logistic regression, $\lambda_n = 0.002369588$.

Chapter 4

Speeding Up Change-points Inference

In this chapter, we design distributed algorithms to speed up change-point inference. In Section 4.1, we describe the problem setup, the challenges with existing approach and our proposed inexact update algorithm. In Section 4.2, we provide several analytical results. In Section 4.3, we test our algorithms using different datasets with various machine learning algorithms.

4.1 Algorithm

In Chapter 2 and Chapter 3, we propose to solve the change-points estimation problem using SGL. However, solving SGL with large volume of data (large n) or high dimensional data (large p) requires huge computation power. This motivates us to utilize distributed learning techniques to speed up change-points inference. Notice that the SGL problem can be written in the form of

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \varphi(\boldsymbol{\beta}) := \mathcal{L}(\boldsymbol{\beta}) + \mathcal{R}(\boldsymbol{\beta}), \tag{4.1}$$

in which $\mathcal{L}(\boldsymbol{\beta})$ is the loss function that measures how well the parameters $\boldsymbol{\beta}$ fit the data, and $\mathcal{R}(\boldsymbol{\beta})$ is the penalty function that measures model complexity.

Notice that (4.1) represents a wide class of machine learning problems including SGL. Different forms of $\mathcal{L}$ and $\mathcal{R}$ lead to different popular machine learning algorithms. The particular choice of $\mathcal{L}$ and $\mathcal{R}$ in (3.1) leads to SGL which is used for change-points inference. Besides the choice of $\mathcal{L}$ and $\mathcal{R}$ in (3.1), there are other forms of $\mathcal{L}$ and $\mathcal{R}$. Other forms of $\mathcal{L}$ and $\mathcal{R}$ are summarized in Table 1.1.

The optimization problem (4.1) has been extensively investigated in the centralized setting where all data is stored in one machine, e.g. in [8,78] and references therein. In this chapter, to speed up the computation, we consider a distributed setup with m machines, each of which stores only part of the dataset. We focus on the challenging feature partition scenario where the whole dataset is partitioned by features. Let $[\mathbf{X}^{[1]}, \mathbf{X}^{[2]}, \cdots, \mathbf{X}^{[m]}]$ be a column-wise partition of the dataset. Machine i stores $\mathbf{X}^{[i]} \in \mathbb{R}^{n \times d_i}$ and hence $\sum_{i=1}^{m} d_i = d$. Let $[\boldsymbol{\beta}^{[1]^T}, \boldsymbol{\beta}^{[2]^T}, \cdots, \boldsymbol{\beta}^{[m]^T}]^T \in \mathbb{R}^d$ be the parameter vector of the corresponding partition with $\boldsymbol{\beta}^{[i]} \in \mathbb{R}^{d_i}$. Let $\mathbf{x}^{[i,j]} \in \mathbb{R}^{d_i}$ and $\mathbf{X}^{[i]} := [\mathbf{x}^{[i,1]}, \mathbf{x}^{[i,2]}, \cdots, \mathbf{x}^{[i,n]}]^T$. Figure 4.1 illustrates the scenario considered along with notation mentioned above.

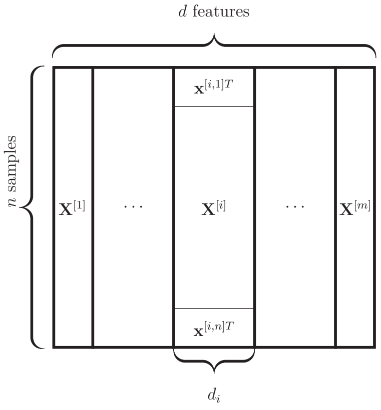

Figure 4.1: Feature partitioned data matrix $\mathbf{X}$ for m machines with $\sum_{i=1}^{m} d_i = d$.

To facilitate the analysis, we make following assumptions that are typically made in the literature, for example see [35, 73, 74].

Assumption 1. *$\mathcal{L}$ is strongly convex and $\mathcal{R}$ is convex.*

Remark 13. *Assumption 1 implies that*

$$\begin{aligned}\mathcal{L}(\boldsymbol{u}) &\geq \mathcal{L}(\boldsymbol{v}) + \langle \nabla\mathcal{L}(\boldsymbol{v}), \boldsymbol{u}-\boldsymbol{v}\rangle + \frac{\mu_{\mathcal{L}}}{2}\|\boldsymbol{u}-\boldsymbol{v}\|^2, &(4.2)\\ \mathcal{R}(\boldsymbol{u}) &\geq \mathcal{R}(\boldsymbol{v}) + \langle \mathrm{s}(\boldsymbol{v}), \boldsymbol{u}-\boldsymbol{v}\rangle, &(4.3)\end{aligned}$$

where $\mathrm{s}(\boldsymbol{v}) \in \partial\mathcal{R}(\boldsymbol{v})$ and $\mu_{\mathcal{L}} > 0$ with $\partial\mathcal{R}(\boldsymbol{v})$ being the subgradient of $\mathcal{R}$ at $\boldsymbol{v}$.

Assumption 2. *The loss function $\mathcal{L}$ is differentiable and there exists a positive semidefinite matrix $\mathbf{M}$ such that*

$$\mathcal{L}(\boldsymbol{\beta}+\mathbf{h}) \leq \mathcal{L}(\boldsymbol{\beta}) + \langle \nabla\mathcal{L}(\boldsymbol{\beta}), \mathbf{h}\rangle + \frac{1}{2}\mathbf{h}^T\mathbf{M}\mathbf{h}. \tag{4.4}$$

Remark 14. *Let the largest eigenvalue of $\mathbf{M}$ be upper bounded by L. Then* (4.4) *implies*

$$\mathcal{L}(\boldsymbol{\beta}+\mathbf{h}) \leq \mathcal{L}(\boldsymbol{\beta}) + \langle \nabla\mathcal{L}(\boldsymbol{\beta}), \mathbf{h}\rangle + \frac{L}{2}\|\mathbf{h}\|^2. \tag{4.5}$$

(4.4) *and* (4.5) *are equivalent since* (4.5) *can be written into form of* (4.4) *by taking $\mathbf{M} = L\mathbf{I}$. This assumption coupled with Assumption 1 implies that the derivative of the loss function is Lipschitz continuous [79, Theorem 2.1.5]:*

$$\|\nabla\mathcal{L}(\boldsymbol{\beta}+\mathbf{h}) - \nabla\mathcal{L}(\boldsymbol{\beta})\| \leq L\|\mathbf{h}\|. \tag{4.6}$$

Remark 15. *Combining* (4.2) *and* (4.4)*, we have $\mu_{\mathcal{L}} \leq L$.*

Assumption 3. *The penalty function is separable for each machine*

$$\mathcal{R}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \mathcal{R}_i(\boldsymbol{\beta}^{[i]}), \tag{4.7}$$

where $\mathcal{R}_i$ is a $\mathbb{R}^{d_i} \to \mathbb{R}$ function.

Under these assumptions, in [35], the authors propose a distributed coordinate descent algorithm to solve problem (4.1). For reader's convinence, we list the algorithm proposed [35] in Algorithm 5 below with modified notation. In this algorithm, we use $\boldsymbol{\beta}_k$ to denote the parameter at kth iteration, $\boldsymbol{\beta}_k^l$ to denote the lth element in vector $\boldsymbol{\beta}_k$, $\nabla\mathcal{L}(\boldsymbol{\beta}_k)_l$ to denote the lth element of the gradient and $\mathbf{M}_{ll}$ to denote the lth element of the diagonal of $\mathbf{M}$.

Algorithm 5 Distributed Coordinate Descent of [35]

```
Input: Step-size parameter $\gamma > 0$; $\tau$ to control the number of coordinates to optimize at each iteration
$k = 0$
while a stopping condition is not satisfied do
    for each machine $i$ in parallel do do
        Pick a random set of coordinates $\hat{S}_i$ of $\mathbf{X}^{[i]}$ with cardinality $|\hat{S}_i| = \tau$
        for each feature index $l$ in $\hat{S}_i$ do
            $\boldsymbol{\beta}_{k+1}^l = \boldsymbol{\beta}_k^l + \text{argmin}_\delta \left[\nabla\mathcal{L}(\boldsymbol{\beta}_k)_l\delta + \frac{\mathbf{M}_{ll}\gamma}{2}\delta^2 + \mathcal{R}_l(\boldsymbol{\beta}_k^l + \delta)\right]$
        end for
    end for
    $k = k + 1$
end while
```

This algorithm depends on the lth element of the gradient $\nabla\mathcal{L}(\boldsymbol{\beta}_k)$. However, it should be noted that the gradient at each iteration cannot be computed locally by each machine.

For example, if the loss function is residual sum of squares, then the gradient at each machine actually involves data at all other machines. In particular, in this case, the loss function is

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2}\left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|^2 = \frac{1}{2}\left\|\mathbf{y} - \sum_{i=1}^{m}\mathbf{X}^{[i]}\boldsymbol{\beta}^{[i]}\right\|^2. \tag{4.8}$$

For machine i^*, its local gradient vector is

$$\nabla\mathcal{L}(\boldsymbol{\beta}^{[i^*]}) = -\mathbf{X}^{[i^*]T}\left(\mathbf{y} - \sum_{i=1}^{m}\mathbf{X}^{[i]}\boldsymbol{\beta}^{[i]}\right). \tag{4.9}$$

In (4.9), the gradient vector for machine i^* is related to not only its local dataset $\mathbf{X}^{[i^*]}$, but also datasets at all other machines. If at each iteration, we update $\boldsymbol{\beta}_k$, then the communication cost is huge since we need to transfer almost the whole dataset at each iteration to compute $\nabla\mathcal{L}(\boldsymbol{\beta}^{[i^*]})$. In particular, each machine needs to sends $\mathbf{X}^{[i]}$ and $\boldsymbol{\beta}^{[i]}$ so that each machine can compute $\nabla\mathcal{L}(\boldsymbol{\beta}^{[i]})$. Another alternative is to update $\sum_{i=1}^{m}\mathbf{X}^{[i]}\boldsymbol{\beta}^{[i]}$ as in [35]. At each iteration, machine i update $\mathbf{X}^{[i]}\boldsymbol{\beta}^{[i]}$ instead of $\boldsymbol{\beta}^{[i]}$. Even if we update $\sum_{i=1}^{m}\mathbf{X}^{[i]}\boldsymbol{\beta}^{[i]}$, the amount of data transmitted at each iteration by each machine is n since $\mathbf{X}^{[i]}\boldsymbol{\beta}^{[i]}$ is a vector of length n. Hence, *the communication cost is too high if we calculate the exact gradient vector at each iteration*. Motivated by this observation, we focus on designing distributed coordinate gradient descent algorithms with low communication overhead.

To reduce the communication costs associated with the distributed coordinate descent algorithms, we design Algorithm 6. In Algorithm 6, we try to reduce the communication cost by only calculating the exact gradient sparsely (we will discuss when to calculate the exact update in the sequel). In the iterations when exact gradients are not computed, we will use the most recently calculated gradient to compute next update. As the result, instead of using the exact gradients $\nabla\mathcal{L}(\boldsymbol{\beta}_k)$ at each iteration, we will use an approximation of the gradient $\nabla\mathcal{L}(\boldsymbol{\beta}_k) + \mathbf{e}_k$, where $\mathbf{e}_k$ is the approximation error vector, to compute next update. Our main idea is to carefully select the time instants at which exact gradient are computed so that the approximation error vectors caused by the sparsely calculated gradients are well controlled and the algorithm still converges.

Now, we provide more details about Algorithm 6. Lines 3-21 are the main body of the algorithm and can be split into two parts. Lines 4-10 form the exact update part and lines 11-20 form the inexact update part. The inputs of the algorithm are the initial starting point $\boldsymbol{\beta}_0$ and a nonnegative sequence $\{\epsilon_k\}$ that will be used as thresholds to determine whether we should perform exact update or inexact update at iteration k.

Line 4 specifies the conditions when we will enter the exact update part. In particular, for the first iteration, i.e., $k = 0$ we perform exact update. For iterations $k \geq 1$, we perform exact updates if the distance between the current parameter vector and the previous parameter vector is large (exceed the threshold ϵ_k). The main intuition is that a large distance between the current parameter vector and the previous parameter vector implies that the objective function is changing fast in the neighborhood of the current parameter vector, and hence we should calculate the exact gradient. Notice in line 4 of Algorithm 6, we need to compute $\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\|$ to judge whether it exceeds the threshold ϵ_k at iteration k. To compute this, each machine i can compute and transmit a scalar $\|\boldsymbol{\beta}_k^{[i]} - \boldsymbol{\beta}_{k-1}^{[i]}\|$ and we have $\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\| = \sum_{i=1}^{m} \|\boldsymbol{\beta}_k^{[i]} - \boldsymbol{\beta}_{k-1}^{[i]}\|$. The amount of communication for each machine is 1 for the step. Furthermore, we should note that the algorithm requires to memorize the $\boldsymbol{\beta}_{k-1} \in \mathbb{R}^d$ or $\boldsymbol{\beta}_{k-1}^{[i]} \in \mathbb{R}^{d_i}$ for each machine i to accomplish this step. Line 5 updates the exact gradient vector which requires communication whose amount of data transferred is n from the example in (4.8) and (4.9). Line 6-10 perform the classic exact update. For an exact iteration

$$\begin{aligned}\boldsymbol{\beta}_k &\overset{(a)}{=} \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}}\langle\nabla\mathcal{L}(\boldsymbol{\beta}), \boldsymbol{\beta}-\boldsymbol{\beta}_{k-1}\rangle + \frac{L}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}_{k-1}\|^2 + \mathcal{R}(\boldsymbol{\beta}) \\ &= \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}}\frac{1}{L}\mathcal{R}(\boldsymbol{\beta}) + \frac{1}{2}\left\|\boldsymbol{\beta} - \left(\boldsymbol{\beta}_{k-1} - \frac{1}{L}\nabla\mathcal{L}(\boldsymbol{\beta}_{k-1})\right)\right\|^2 \\ &\overset{(b)}{=} prox_{\frac{1}{L}\mathcal{R}}\left(\boldsymbol{\beta}_{k-1} - \frac{1}{L}\nabla\mathcal{L}(\boldsymbol{\beta}_{k-1})\right),\end{aligned}$$

where (a) is based on $\sum_{i=1}^m\langle\nabla\mathcal{L}(\boldsymbol{\beta}^{[i]}), \boldsymbol{\beta}^{[i]}-\boldsymbol{\beta}_{k-1}^{[i]}\rangle = \langle\nabla\mathcal{L}(\boldsymbol{\beta}), \boldsymbol{\beta}-\boldsymbol{\beta}_{k-1}\rangle$ and Assumption 3, and in (b) we use the definition of proximity operator $prox_f(\boldsymbol{v}) = \operatorname{argmin}_{\mathbf{w}} f(\mathbf{w}) + \frac{1}{2}\|\mathbf{w}-\boldsymbol{v}\|^2$.

As illustrated in Line 11, for $k \geq 1$, if $\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\| \leq \epsilon_k$, then we enter the inexact update part. The main idea is that a small distance between the current parameter vector and the previous parameter vector implies that the value of the objective function do not change dramatically in the neighborhood of the current parameter vector, and hence we can use previously calculated gradient to compute the next update. In line 12, we take the $\boldsymbol{\beta}_{fixed} = \boldsymbol{\beta}_{k-1}$ to utilize the condition in line 11 that $\boldsymbol{\beta}_k$ is very close to $\boldsymbol{\beta}_{k-1}$. Combining the condition for the while loop in line 13 and the fact that this is the first step to enter the while loop, we know that $(k-1)$th iteration is the exact iteration, which means $\nabla\mathcal{L}(\boldsymbol{\beta}_{fixed})$ is already computed in the exact update part. It should be noticed that $\nabla\mathcal{L}(\boldsymbol{\beta}_{fixed})$ should be stored in machines. To accomplish this, each machine i can store $\nabla\mathcal{L}(\boldsymbol{\beta}_{k-1}^{[i]}) \in \mathbb{R}^{d_i}$. In line 13-18, we continuously use $\nabla\mathcal{L}(\boldsymbol{\beta}_{fixed})$ as the approximation gradient vector instead of computing the exact one until it can no longer be used. Lines 14-17 perform update as in line 6-10. The only difference here is that the approximated gradient vector $\nabla\mathcal{L}(\boldsymbol{\beta}_{fixed})$ is used. In these inexact iterations, we have

$$\begin{aligned}\boldsymbol{\beta}_k &= prox_{\frac{1}{L}\mathcal{R}}\left(\boldsymbol{\beta}_{k-1} - \frac{1}{L}\nabla\mathcal{L}(\boldsymbol{\beta}_{fixed})\right)\\ &= prox_{\frac{1}{L}\mathcal{R}}\left(\boldsymbol{\beta}_{k-1} - \frac{1}{L}\nabla\mathcal{L}(\boldsymbol{\beta}_{k-1}) + \mathbf{e}_{k-1}\right),\end{aligned}$$

in which $\mathbf{e}_{k-1} = \frac{1}{L}\left(\nabla\mathcal{L}(\boldsymbol{\beta}_{k-1}) - \nabla\mathcal{L}(\boldsymbol{\beta}_{fixed})\right)$. So we actually perform a proximal gradient method at each step, the only difference is that we introduce error in inexact iterations.

We will discuss in detail how to select the sequence $\{\epsilon_k\}$ in Section 4.2. Informally, to guarantee the convergence of the algorithm, we choose $\{\epsilon_k\}$ to be summable. Furthermore, we should select the sequence $\{\|\epsilon_k\|\}$ to be diminishing as well. This selection ensures that, as we get closer to the optimal solution, the error introduced by the inexact update also gets smaller. Thus, this selection will prevent injecting a large error into the gradient vector when we get to the close neighborhood of the optimal solution, as a large gradient error will lead to large deviation in the result that would make it difficult for the algorithm to converge.

Algorithm 6 Distributed Coordinate Descent Algorithm with Inexact Update

```
input an initial point $\boldsymbol{\beta}_0$ and a nonnegative sequence $\{\epsilon_k\}$.
$k = 0.$
while a stopping condition is not satisfied do
    if $k == 0$ or $\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\| > \epsilon_k$ then    ▷ need communication, amount of $m$
        compute exact value of $\nabla\mathcal{L}(\boldsymbol{\beta}_k)$    ▷ need communication
        for each machine $i$ in parallel do do
            $\mathbf{h}^{[i]*} = \text{argmin}_{\mathbf{h}^{[i]}\in\mathbb{R}^{d_i}} \langle\nabla\mathcal{L}(\boldsymbol{\beta}_k^{[i]}), \mathbf{h}^{[i]}\rangle + \frac{L}{2}\|\mathbf{h}^{[i]}\|^2 + \mathcal{R}_i(\boldsymbol{\beta}_k^{[i]} + \mathbf{h}^{[i]}).$    ▷ update using exact gradient
            $\boldsymbol{\beta}_{k+1}^{[i]} = \boldsymbol{\beta}_k^{[i]} + \mathbf{h}^{[i]*}$
        end for
        $k = k + 1$
    else
        $\boldsymbol{\beta}_{fixed} = \boldsymbol{\beta}_{k-1}$    ▷ $\nabla\mathcal{L}(\boldsymbol{\beta}_{k-1})$ is known, no communication
        while stopping condition is not satisfied and $\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{fixed}\| \le \epsilon_k$ do
            for each machine $i$ in parallel do do
                $\mathbf{h}^{[i]*} = \text{argmin}_{\mathbf{h}^{[i]}\in\mathbb{R}^{d_i}} \langle\nabla\mathcal{L}(\boldsymbol{\beta}_{fixed}^{[i]}), \mathbf{h}^{[i]}\rangle + \frac{L}{2}\|\mathbf{h}^{[i]}\|^2 + \mathcal{R}_i(\boldsymbol{\beta}_k^{[i]} + \mathbf{h}^{[i]}).$    ▷ update using inexact gradient
                $\boldsymbol{\beta}_{k+1}^{[i]} = \boldsymbol{\beta}_k^{[i]} + \mathbf{h}^{[i]*}$
            end for
            $k = k + 1$
        end while
    end if
end while
```

4.2 Performance Analysis

In this section, we analyze the convergence rate and communication cost of Algorithm 6.

First, we show two propositions about the general setting of threshold sequence $\{\epsilon_k\}$ to perform inexact updates.

Proposition 19. *If we update using some approximate gradient $\nabla\mathcal{L}(\boldsymbol{\beta}_{fixed})$ instead of exact gradient $\nabla\mathcal{L}(\boldsymbol{\beta}_k)$, then we have*

1. *If $\frac{\|\boldsymbol{\beta}_k-\boldsymbol{\beta}_{fixed}\|}{\|\boldsymbol{\beta}_{k+1}-\boldsymbol{\beta}_k\|} \leq \frac{1}{2}$, we have $\varphi(\boldsymbol{\beta}_{k+1}) - \varphi(\boldsymbol{\beta}_k) \leq 0$;*

2. *If $\frac{\|\boldsymbol{\beta}_k-\boldsymbol{\beta}_{fixed}\|}{\|\boldsymbol{\beta}_{k+1}-\boldsymbol{\beta}_k\|} \leq \frac{\mu_{\mathcal{L}}}{5L}$, we have*

$$\begin{aligned}&\varphi(\boldsymbol{\beta}_{k+1}) - \varphi(\boldsymbol{\beta}^*) + \frac{1}{2}\left(L - \frac{3}{5}\mu_{\mathcal{L}}\right)\left\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}^*\right\|^2 \\ &\leq \left(\frac{5L-4\mu_{\mathcal{L}}}{5L-3\mu_{\mathcal{L}}}\right)\left(\frac{1}{2}\left(L - \frac{3}{5}\mu_{\mathcal{L}}\right)\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|^2\right). \qquad (4.10)\end{aligned}$$

Proof. Please see Appendix A.12. □

After introducing inexactness in the gradient used to calculate the next update, generally, there is no guarantee for the convergence of algorithm if the inexactness is not carefully selected or controlled. Proposition 19 provides metrics that can be used in the algorithm to guarantee that the value of the objective function converges to the minimum. In particular, condition 1) of Proposition 19 gives a sufficient condition for the value of the objective function to be non-increasing for two consecutive iterations. The condition 2) of Proposition 19 is strict if $\frac{\mu_{\mathcal{L}}}{5L} << 1$. However, it provides an exponential convergence rate which is much larger than 1) provides. To check the condition for the conditions, we need to compute the norms $\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{fixed}\|$ and $\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|$ which requires communication of a scalar for each machine and is much less than n for computing exact gradient vector. Proposition 19 provides a way for us to judge if the current inexact iteration is good or not. If we are more conservative, after each inexact update, we can check whether or not the conditions in Proposition 19 hold. If none of the conditions holds, we can give up the results of the current inexact iteration and go back to the exact iteration.

Next, we give an explicit value sequence of $\{\epsilon_k\}$ and provide two propositions about the convergence rate and an upper bound on the number of iterations at which exact update is carried out. In the following, we analyze the basic version of our algorithm, namely at each step we do not additionally use conditions in Proposition 1 to check whether the current inexact update is good enough or not.

Proposition 20. *Let D be an upper bound of $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|$. If we set $\epsilon_k = \frac{\mu_0}{L}\left(1 - \frac{\mu_{\mathcal{L}} - \mu_0}{L}\right)^k D$, where $0 < \mu_0 < \mu_{\mathcal{L}}$, then*

$$\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\| \leq \left(1 - \frac{\mu_{\mathcal{L}} - \mu_0}{L}\right)^k D. \tag{4.11}$$

Proof. Please see Appendix A.13. □

Proposition 20 shows that if we choose $\{\epsilon_k\}$ in this way, we can still achieve an exponential convergence rate even with inexact gradient updates, although the convergence speed is slower than the algorithm without inexactness. The convergence rate also depends on the estimated upper bound D and we want D to be as close to $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|$ as possible.

Proposition 21. *Let N be the number of iterations until convergence for Algorithm 6. The number of exact update iterations (i.e., the iterations when large communication overhead is needed) is at most*

$$N \frac{\log\left(1 - \frac{\mu_{\mathcal{L}} - \mu_0}{L}\right)}{\log\left(1 - \frac{\mu_{\mathcal{L}}}{L}\right)}.$$

Proof. Please see Appendix A.14. □

Proposition 21 provides an upperbound on the number of iterations where exact updates are carried out. If inexact updates are not introduced, the total number of iterations k to achieve $\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\| \leq \epsilon$ is $\frac{\log \epsilon - \log \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|}{\log\left(1-\frac{\mu_{\mathcal{L}}}{L}\right)}$. If inexact updates are introduced, the corresponding number of iterations is $\frac{\log \epsilon - \log D}{\log\left(1-\frac{\mu_{\mathcal{L}}-\mu_0}{L}\right)}$. Notice if D is a tight bound and we use Proposition 21, then we find that the communication iterations are approximately equal. However, this is a comparison of worst cases of these two methods. In practice, we find that by introducing inexact updates is more communication efficient on average, which is shown in the next section.

A natural question would arise: what if we cannot find a upper bound D? To answer this question, we notice that $\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|$ converges to 0 as $k \to \infty$ without inexactness. If we cannot find a upper bound D, then a lower bound is also fine with this method. Since we can use exact iterations for first i rounds to make $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}^*\| \leq D$. Then we can treat this $\boldsymbol{\beta}_i$ as the initial point $\boldsymbol{\beta}_0$.

4.3 Numerical Examples

In this section, we provide numerical examples to illustrate our results using synthesized data and real data. We will preform our tests for different popular machine learning algorithms: Lasso, SVM and logistic regression. For SVM, we use the smoothed hinge loss. Furthermore, for SVM and logistic regression, we choose ℓ_1 norm as the penalty function. Hence, throughout this section, the penalty function is $\mathcal{R}(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$ with $\lambda = 0.001$.

We compare our algorithm with the case in which the exact update is calculated at every iteration, which is equivalent to set $\epsilon_k = 0$ for all k in Algorithm 6. For each of Lasso, SVM and logistic regression, we test Algorithm 6 using both synthesized data and real data respectively.

For synthesized data simulation, we use the error sequence $\{\epsilon_k\}$ stated in Proposition 20. The data matrix $\mathbf{X}$ is generated randomly with fixed known maximal and minimal eigenvalues of $\mathbf{X}^T\mathbf{X}$. The vector $\mathbf{y}$ is generated by linear regression for Lasso, and $\mathbf{y}$ consists of class labels $(\{+1, -1\})$ for SVM and logistic regression. For Lasso and SVM with the smoothed hinge loss, (4.4) holds for their loss functions with $\mathbf{M} = \mathbf{X}^T\mathbf{X}$. For logistic regression, (4.4) holds with $\mathbf{M} = \frac{1}{4}\mathbf{X}^T\mathbf{X}$. Let $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the maximal and minimal eigenvalue of $\mathbf{M}$ respectively. So for Lasso and SVM with smoothed hinge loss, $L = 2\lambda_{\max}(\mathbf{X}^T\mathbf{X})$ and $\mu_{\mathcal{L}} = 2\lambda_{\min}(\mathbf{X}^T\mathbf{X})$. For logistic regression, $L = \frac{1}{2}\lambda_{\max}(\mathbf{X}^T\mathbf{X})$ and $\mu_{\mathcal{L}} = \frac{1}{2}\lambda_{\min}(\mathbf{X}^T\mathbf{X})$. We set $L = 20000$ and $\mu_{\mathcal{L}} = 2$, and we generate $\mathbf{X}$ with corresponding fixed $\lambda_{\max}(\mathbf{X}^T\mathbf{X})$ and $\lambda_{\min}(\mathbf{X}^T\mathbf{X})$. For synthesized data, we set the number of samples $n = 2000$ and the number of features d from 10 to 400 with increments being 10.

For simulations with real datasets, we use various datasets that are publicly available. Since the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are unknown and hard to compute for real datasets, the error sequence in Proposition 20 can no longer be used. For practical concerns, we simply set $\epsilon_k = (1-\alpha)^\tau D$, where τ is the number of inexact iterations so far. Here we set L to be an easily computable value $\frac{L}{2} = \|\mathbf{X}\|_F^2 \geq \|\mathbf{X}^T\mathbf{X}\|_F \geq \lambda_{\max}(\mathbf{X}^T\mathbf{X})$, where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. We run simulations for different values of α to show the performance of this error sequence in practice.

We run simulations in pseudo-distributed environment. In our simulations, we care about the inexact communication iterations, which do not depend on the number of machines and the ways to partition the features of the dataset. Noticing that no matter how many machines we have, the inexact communication iterations are the same; therefore we run Algorithm 6 in one machine to simulate the case in a distributed cluster of machines.

4.3.1 Lasso

We first conduct our algorithm to solve Lasso, where $\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ and $\mathcal{R}(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$ with $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{\beta} \in \mathbb{R}^d$.

Synthesized data

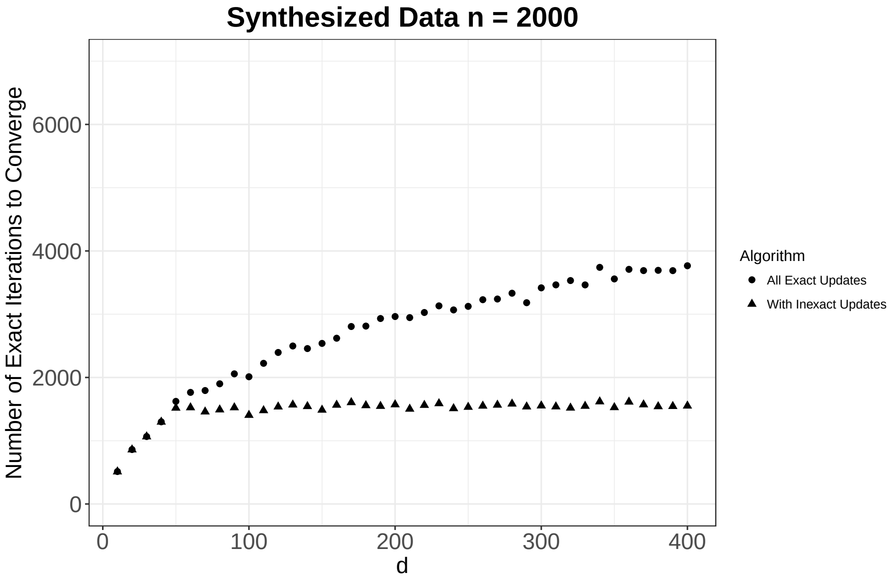

Figure 4.2: Number of exact communication iterations for Lasso.

In our simulation, we compare Algorithm 6 with that of Algorithm 6 without inexactness (Algorithm 6 without Lines 11 - 18). Comparing Figure 4.2 and Figure 4.3, we find that the values of the objective function after minimization by two methods are very close. However, the proposed scheme reduces nearly half of the total communications.

Real data

Next, we test our algorithm on real dataset: Communities and Crime Unnormalized Data Set [3]. The dataset contains statistics related to crime and social economics from 1990 US Census, 1990 US LEMAS survey and 1995 FBI UCR. The features contain statistics

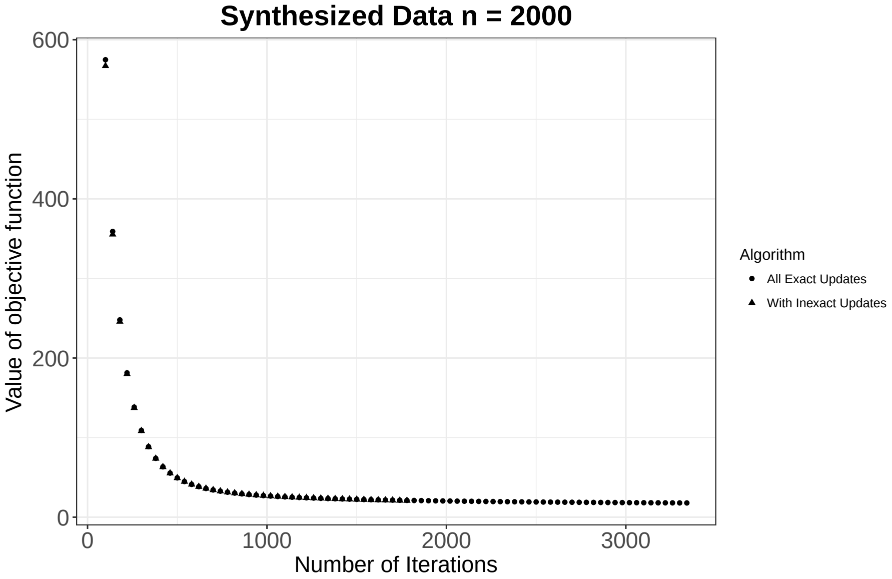

Figure 4.3: Values of the objective function versus the number of iterations for Lasso $d = 400$.

such as community population, per capita income, police operating budget and violent crime rate etc.

In this example, we study the murder rate (per 100K people) and try to build a sparse linear regression model between the murder rate and all other numeric variable in the dataset. Here we have $n = 2215$ and $d = 103$ (we omit the data features with missing data).

The results computed by the algorithm without inexactness (with $\{\epsilon_k = 0\}$) is listed in Table 4.1. The results with inexactness are summarized in Table 4.2.

Table 4.1: Basic algorithm for crime data

Iterations	Value of objective function
15300	681.4654

Table 4.2: Algorithm with inexact iterations for crime data

D	α	Total iterations	Inexact iterations	Value of objective function
1	10^{-1}	15299	95	681.4653
1	10^{-2}	15287	992	681.4674
1	10^{-3}	13735	9965	681.6371

Table 4.2 shows that our scheme does not work well for $\alpha = 10^{-1}$ and works well for $\alpha = 10^{-3}$. Although this thesis does not provide theoretical results for the cases where $\mu_{\mathcal{L}}$ is unknown or $\mu_{\mathcal{L}} = 0$, Table 4.2 shows that our scheme works for these case in practice. It would be interesting to extend to these two cases in the theoretical analysis.

4.3.2 SVM

Next, we study a classification task and perform SVM. Here we use smoothed hinge loss so that the loss function is differentiable as required in Assumption 2:

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{j=1}^{n} \max\{0, 1 - y_j \mathbf{x}_j^T \boldsymbol{\beta}\}^2.$$

Let $\nabla\mathcal{L}_j(\boldsymbol{\beta}) := \max\{0, 1 - y_j \mathbf{x}_j^T \boldsymbol{\beta}\}^2$. We have

$$\nabla\mathcal{L}_j(\boldsymbol{\beta}) = \begin{cases} \mathbf{0}, & 1 - y_j \mathbf{x}_j^T \boldsymbol{\beta} \le 0 \\ -2y_j(1 - y_j \mathbf{x}_j^T \boldsymbol{\beta})\mathbf{x}_j, & 1 - y_j \mathbf{x}_j^T \boldsymbol{\beta} > 0 \end{cases}.$$

So

$$\nabla\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{j=1}^{n} \nabla\mathcal{L}_j(\boldsymbol{\beta}).$$

We set $\mathcal{R}(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$.

Synthesized Data

The data matrix $\mathbf{X}$ is generated as described in Section 4.3.1. We randomly generate a coefficient vector $\boldsymbol{\beta}$. We label y_j as $+1$ if $\boldsymbol{\beta}^T \mathbf{x}_j > 0$ and -1 otherwise. These components form the label vector $\mathbf{y} = [y_1, \ldots, y_n]$.

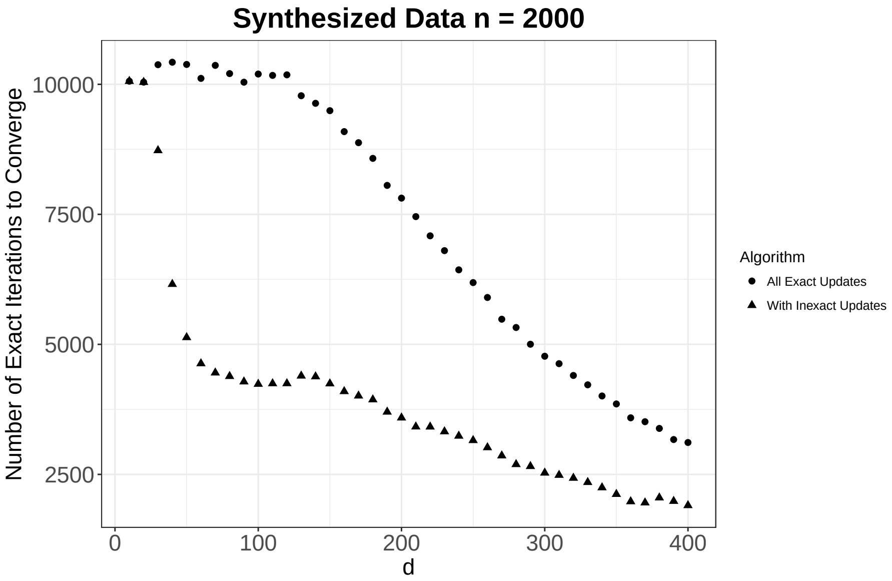

Figure 4.4: Number of exact communication iterations for SVM.

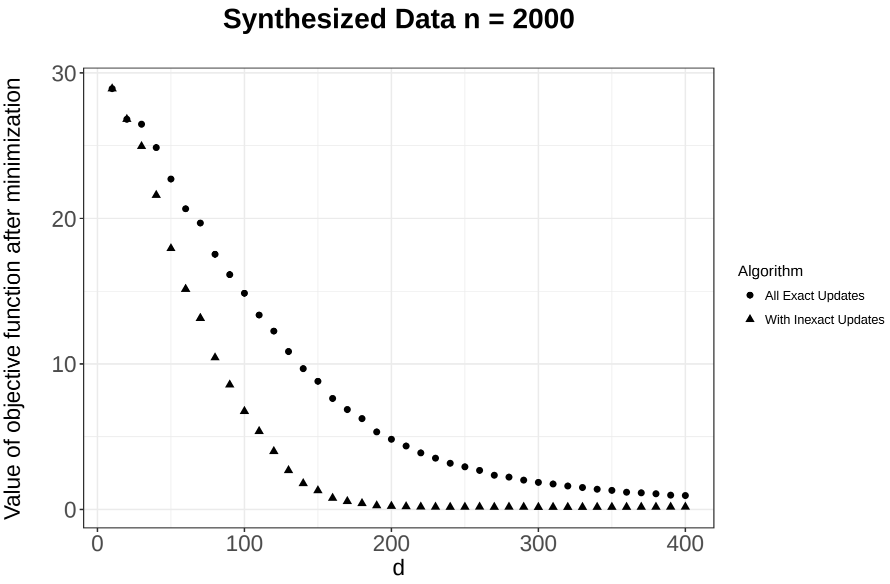

Figure 4.5: Objective function value after minimization for SVM.

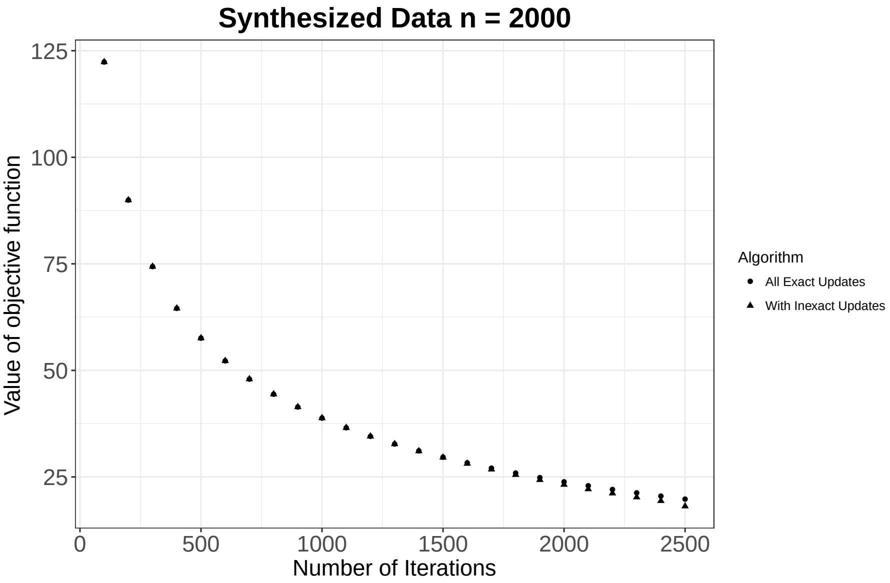

Figure 4.6: Value of the objective function with the number of dimension $d = 200$ for SVM.

Figure 4.4 and Figure 4.5 show that the algorithm with inexactness outperforms the benchmark. For Figure 4.6, the objective function is recorded every 100 iterations. The exact method needs 2500 exact iterations while the inexact method needs 1270 exact iterations. It is worth mentioning that in some other simulations we run, we notice the objective function value is increasing for some iterations. This is due to ϵ_k is large hence the error introduced is too large for these iterations. Since ϵ_k is strictly decreasing, after enough number of iterations, the value of objective function jumps back abruptly to normal. In total, it outperforms the benchmark.

Real data

The dataset we use here is $a9a$ [2]. This dataset is also known as "Census Income" dataset. It contains information of whether income exceeds $50K/year$ based on census data. Features in the dataset include each person's age, work type, education and capital-gain etc. The dataset has $n = 22696$ samples and its dimension is $d = 123$. Our task here is to classify data points with income greater than and less than $50K/year$.

Table 4.3: Basic algorithm for $a9a$

Iterations	Value of objective function
8773	6880.909

Table 4.4: Algorithm with inexact iterations for $a9a$

D	α	Total iterations	Inexact iterations	Value of objective function
1	10^{-1}	8762	95	6880.902
1	10^{-2}	8232	992	6880.685
1	10^{-3}	22915	9960	6876.830

Different from the results obtained in the Communities and Crime Data Set example, this comparison implies that α does not have a direct linear relationship with the actual performance.

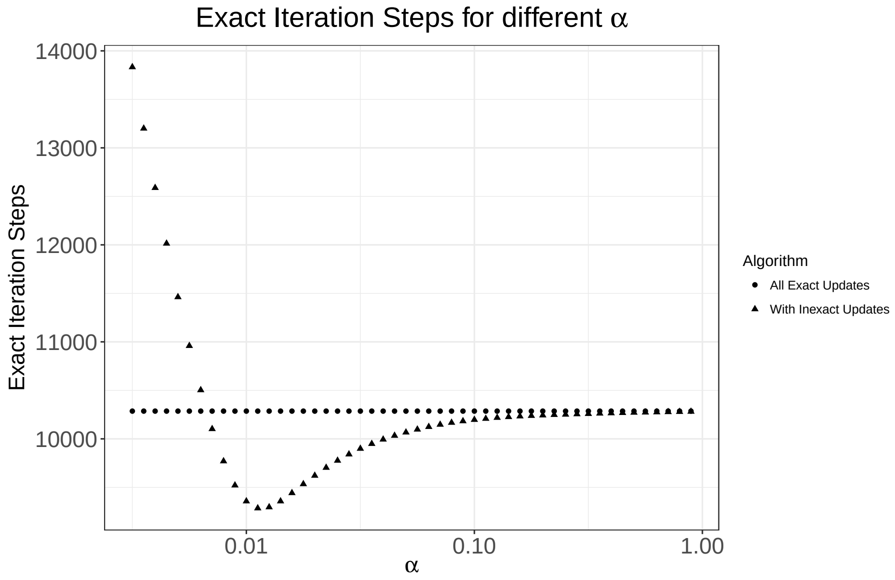

Figure 4.7: Number of iterations comparison for different values of α

Figure 4.7 further illustrates the impact of different α and it shows the best performance is obtained when α is around 10^{-2}. Figure 4.7 further shows that we would rather choose a larger value for α. If α is large, then the error threshold decrease more quickly, so it at least can guarantee its performance similar algorithm without inexactness. If α is too small, the errors introduced would be too large. In practice, a large number of exact iterations to correct those errors are needed before the stopping condition is met.

It is worth mentioning that the number of exact iterations near $\alpha = 10^{-3}$ of algorithm with inexactness is large does not contradict Proposition 20 and Proposition 21 since here we use a different sequence $\{\epsilon_k\}$ for practical concerns as described at the beginning of this section.

4.3.3 Logistic Regression

Next, we use logistic regression to accomplish a classification task.

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{j=1}^{n} \log(1 + \exp(-y_j \boldsymbol{\beta}^T \mathbf{x}_j)).$$

We have

$$\nabla \mathcal{L}(\boldsymbol{\beta}) = \sum_{j=1}^{n} -\frac{y_j}{1 + \exp(y_j \boldsymbol{\beta}^T \mathbf{x}_j)} \mathbf{x}_j.$$

Synthesized Data

Here we generate the data the same way as described in Section 4.3.2. Here we try to build logistic regression between the data matrix $\mathbf{X}$ and the labels $\mathbf{y}$.

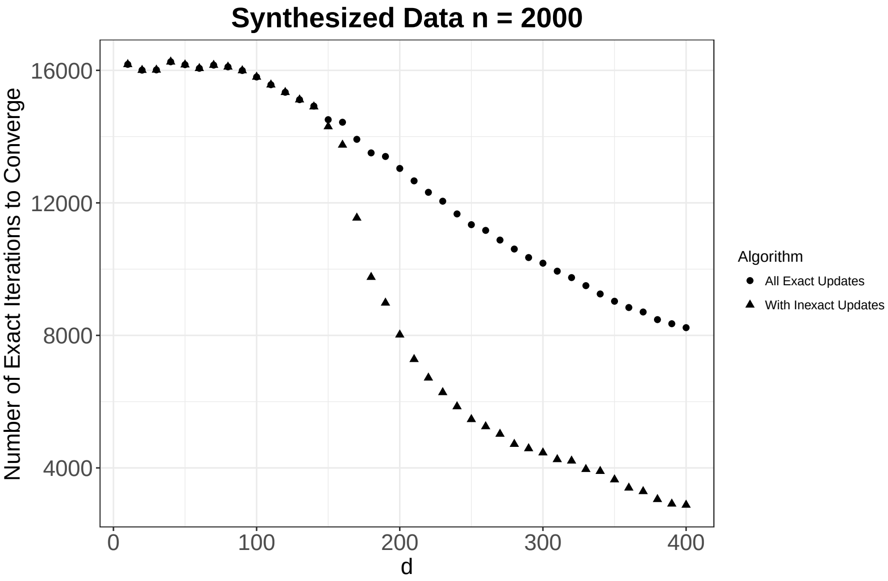

Figure 4.8: Number of exact communication iterations.

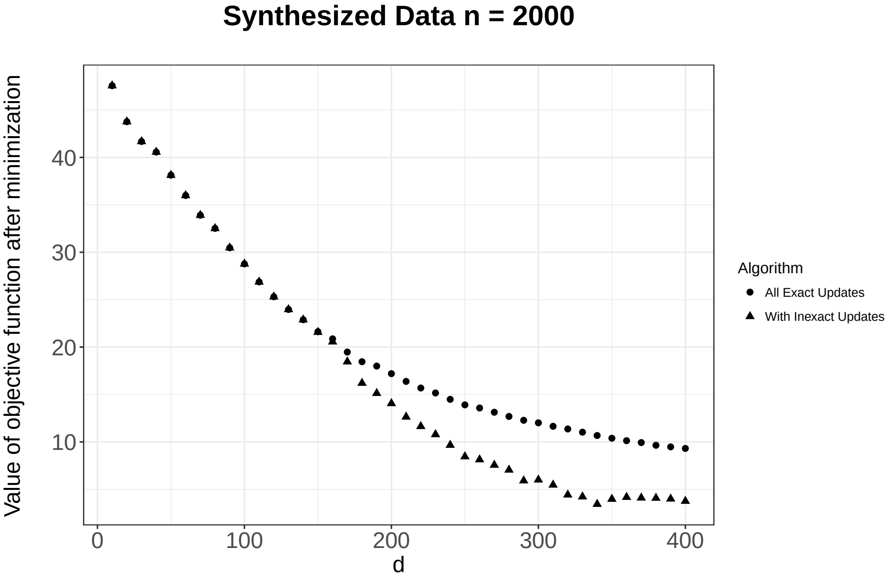

Figure 4.9: Objective function value after minimization.

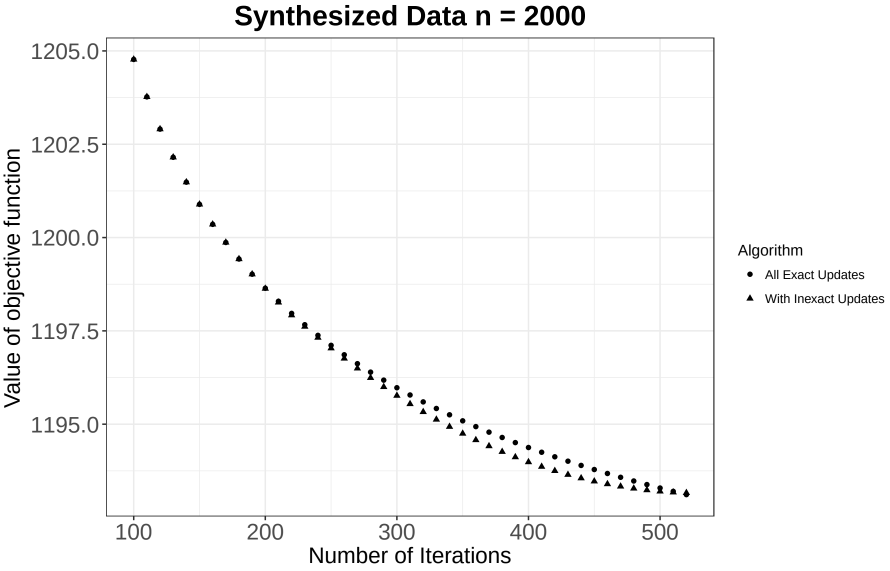

Figure 4.10: Value of the objective function with the number of dimension $d = 200$.

Figure 4.8 and Figure 4.9 show that the algorithm with inexactness outperforms the benchmark. For Figure 4.10, the objective function is recorded every 10 iterations. The exact method needs 560 exact iterations while the inexact method needs 150 exact iterations.

Real Data

Here we use the news20.binary dataset [2] with $n = 19996$ and $d = 1355191$. It is a two-class dataset which consists of news from 20 groups. Each data sample corresponds to one message taken from the news groups. Each data sample is a vector with binary term frequency and then normalized to unit length. The positive class consists of the 10 news groups, and the negative class consists of the other 10 news groups. Here we try to build logistic regression with ℓ_1 penalty for this dataset.

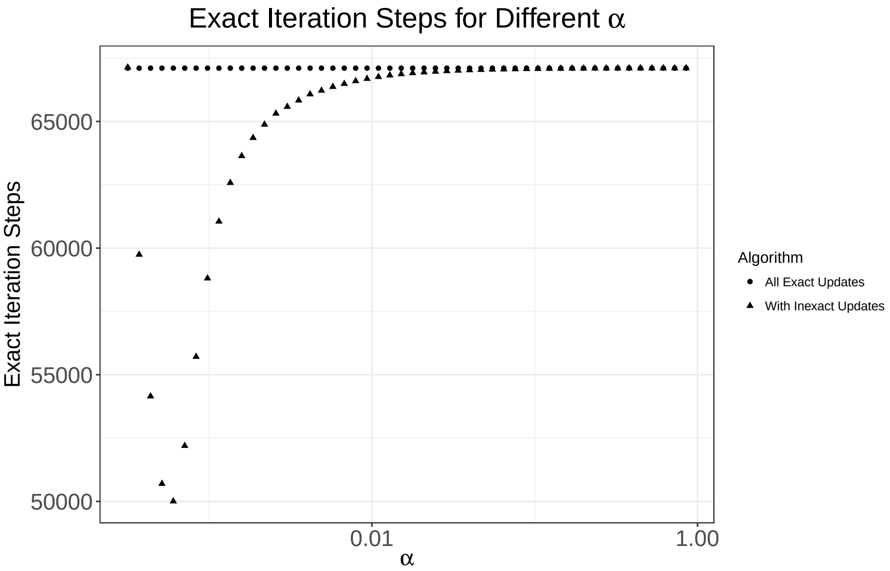

Figure 4.11: Number of exact communication iterations.

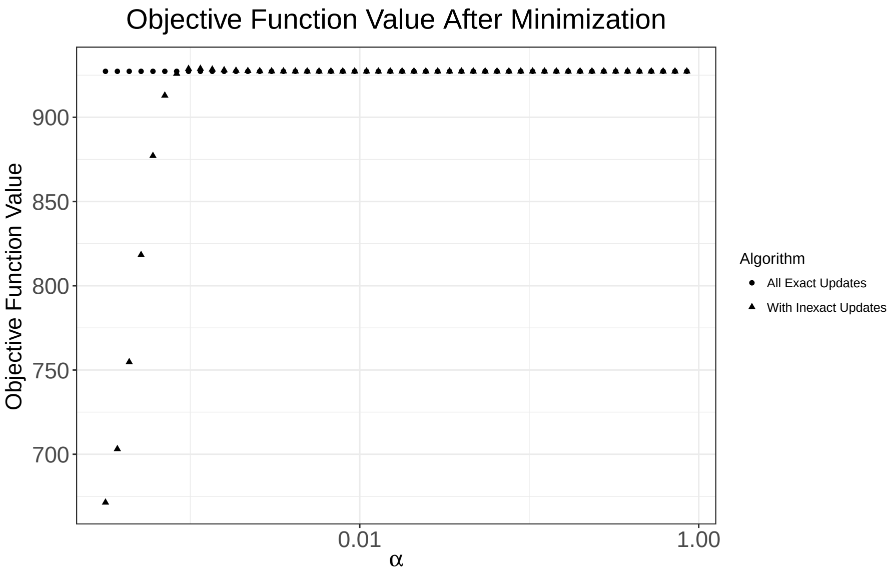

Figure 4.12: Objective function value after minimization.

Figure 4.11 shows the number of exact iterations for different values of α. Figure 4.12 shows the values of objective function at termination.

Chapter 5

Conclusion

This thesis has addressed the need for statistical inference in heterogeneous models. The main contributions are transforming change-points estimation problem into an SGL problem, providing analytical results and designing efficient algorithm empowered by distributed learning.

The method presented and analyzed in Chapter 2 is for low dimensional models. We have shown how to transform the high dimensional multiple change-points problem into a SGL based problem. We have extended our results to high dimensional models in Chapter 3. In Chapter 3, we have also shown that our method can be extended to GLM, which reveals our approach is applicable for a wide range of machine learning algorithms. In both low dimensional and high dimensional settings, we have provided analytical results to show our estimator is consistent. In both cases, we have provided numerical examples to illustrate the analytical results obtained.

To further speed up change-points estimation using SGL, we have proposed a general communication efficient scheme for distributed learning problem of feature partitioned data, which can be used for SGL or other similar tasks. We have proposed an explicit algorithm using inexact updates. We have shown analytical results of the algorithm which reveals its desirable properties under mild assumptions. We further make worst case comparisons. We have shown the efficiency of our proposed algorithm using numerical simulation results.

Appendix A

Proof Details

A.1 Supporting Lemmas

In this section, we provide several supporting lemmas that will be repeatedly used in the proof.

A.1.1 Lemma 4

First we state a lemma that will be used repeatedly in the proofs.

Lemma 4. *Let* $\{\hat{\boldsymbol{\theta}}_t, t = 1, 2, ..., n\}$ *be the solution of the problem* (2.7)*, then*

$$\frac{1}{n}\left\|\sum_{i=t}^{n} \mathbf{x}_i(y_i - \mathbf{x}_i^T \sum_{j=1}^{i} \hat{\boldsymbol{\theta}}_j)\right\|_2 \leq \frac{\lambda_n}{2}\left(\gamma + (1-\gamma)\sqrt{p}\right), \ \forall t. \tag{A.1}$$

Proof. We use KKT condition to prove Lemma 4. To use this condition, we first compute the gradient of $f(\boldsymbol{\theta}) \triangleq -\frac{1}{2}\|\mathbf{Y} - \mathbf{X}\tilde{\mathbf{A}}\boldsymbol{\theta}\|_2^2$ with respect to $\boldsymbol{\theta}$ and obtain

$$\begin{aligned}
\nabla f(\boldsymbol{\theta}) &= \tilde{\mathbf{A}}^T\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{A}}\boldsymbol{\theta}) \\
&= \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \\ & \mathbf{x}_2 & \cdots & \mathbf{x}_n \\ & & \cdots & \mathbf{x}_n \\ & & & \mathbf{x}_n \end{pmatrix} \begin{pmatrix} y_1 - \mathbf{x}_1^T \sum_{j=1}^{1} \boldsymbol{\theta}_j \\ y_2 - \mathbf{x}_2^T \sum_{j=1}^{2} \boldsymbol{\theta}_j \\ \cdots \\ y_n - \mathbf{x}_n^T \sum_{j=1}^{n} \boldsymbol{\theta}_j \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^{n} \mathbf{x}_i(y_i - \mathbf{x}_i^T \sum_{j=1}^{i} \boldsymbol{\theta}_j) \\ \sum_{i=2}^{n} \mathbf{x}_i(y_i - \mathbf{x}_i^T \sum_{j=1}^{i} \boldsymbol{\theta}_j) \\ \cdots \\ \sum_{i=n}^{n} \mathbf{x}_i(y_i - \mathbf{x}_i^T \sum_{j=1}^{i} \boldsymbol{\theta}_j) \end{pmatrix}.
\end{aligned}$$

Since $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^T, \cdots, \boldsymbol{\theta}_t^T, \cdots, \boldsymbol{\theta}_n^T\right)^T$, we can also write

$$\nabla f(\boldsymbol{\theta}) = \left(\left(\frac{\partial f}{\partial \boldsymbol{\theta}_1}\right)^T, \cdots, \left(\frac{\partial f}{\partial \boldsymbol{\theta}_t}\right)^T, \cdots, \left(\frac{\partial f}{\partial \boldsymbol{\theta}_n}\right)^T\right)^T$$

As the result, we have

$$\frac{\partial f}{\partial \boldsymbol{\theta}_t} = \sum_{i=t}^{n} \mathbf{x}_i(y_i - \mathbf{x}_i^T \sum_{j=1}^{i} \boldsymbol{\theta}_j). \tag{A.2}$$

Let $\{\hat{\boldsymbol{\theta}}_t, t = 1, 2, \cdots, n\}$ be the solution of the problem (2.7). Using KKT conditions and (A.2), we have for each $t = 1, \cdots, n$

$$-\frac{2}{n}\sum_{i=t}^{n} \mathbf{x}_i \left(y_i - \mathbf{x}_i^T \sum_{j=1}^{i} \hat{\boldsymbol{\theta}}_j\right) + \gamma\lambda_n \mathbf{u}_t + (1-\gamma)\lambda_n \mathbf{v}_t = \mathbf{0}_{p\times 1}, \tag{A.3}$$

where $\mathbf{u}_t$ and $\mathbf{v}_t$ are subdifferentials of $\sum_{i=1}^{n} \|\hat{\boldsymbol{\theta}}_i\|_2$ and $\sum_{i=1}^{n} \|\hat{\boldsymbol{\theta}}_i\|_1$ at $\hat{\boldsymbol{\theta}}_t$ respectively. Note that we use subdifferentials here since $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_2$ do not have derivatives at $\mathbf{x} = \mathbf{0}_{p\times 1}$. The optimal condition is that zero vector belongs to the subdifferentials.

For $\mathbf{u}_t$ we have,

$$\begin{aligned} \mathbf{u}_t &= \frac{\hat{\boldsymbol{\theta}}_t}{\|\hat{\boldsymbol{\theta}}_t\|_2}, \text{ for } \|\hat{\boldsymbol{\theta}}_t\|_2 \neq 0, \\ \|\mathbf{u}_t\|_2 &\leq 1, \text{ for } \|\hat{\boldsymbol{\theta}}_t\|_2 = 0. \end{aligned}$$

For $\mathbf{v_t}$ we have,

$$\begin{aligned} & v_{t,m} = \text{sign}(\hat{\theta}_{t,m}), \text{ for } 1 \leq m \leq p \text{ and } \hat{\theta}_{t,m} \neq 0, \\ & |v_{t,m}| \leq 1, \text{ for } 1 \leq m \leq p \text{ and } \hat{\theta}_{t,m} = 0, \end{aligned}$$

where $v_{t,m}$ is the mth element of $\mathbf{v_t}$ and $\hat{\theta}_{t,m}$ is the mth element of $\hat{\boldsymbol{\theta}}_t$.

From (A.3), we have

$$\begin{aligned} & \frac{1}{n}\left\|\sum_{i=t}^{n} \mathbf{x}_i \left(y_i - \mathbf{x}_i^T \sum_{j=1}^{i} \hat{\boldsymbol{\theta}}_j\right)\right\|_2 \\ = \; & \frac{\lambda_n}{2}\|\gamma \mathbf{u}_t + (1-\gamma)\mathbf{v}_t\|_2 \\ \leq \; & \frac{\lambda_n}{2}\left(\gamma + (1-\gamma)\sqrt{p}\right). \end{aligned} \tag{A.4}$$

□

Remark 16. *As* $\sum_{j=1}^{i} \hat{\boldsymbol{\theta}}_j = \hat{\boldsymbol{\beta}}_i$*, Lemma 4 can be rewritten as the following equivalent form:*

$$\frac{1}{n}\left\|\sum_{i=t}^{n} \mathbf{x}_i \left(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_i\right)\right\|_2 \leq \frac{\lambda_n}{2}\left(\gamma + (1-\gamma)\sqrt{p}\right), \ \forall t. \tag{A.5}$$

A.1.2 Lemma 5

Lemma 5. *If A3 holds, then*

$$\max_{\substack{1\le r<s\le n+1\\ s-r\ge n\delta_n}} \left\| \frac{1}{s-r}\sum_{t=r}^{s-1} \mathbf{x}_t\epsilon_t \right\|_2 = o_P(J_{min}). \tag{A.6}$$

Proof. For a given constant $c>0$,

$$\begin{aligned} & P\left(\max_{\substack{1\le r<s\le n+1\\ s-r\ge n\delta_n}} \left\| \frac{1}{s-r}\sum_{t=r}^{s-1} \mathbf{x}_t\epsilon_t \right\|_2 \ge cJ_{min} \right) \\ \le & \sum_{\substack{1\le r<s\le n+1\\ s-r\ge n\delta_n}} P\left(\left\| \frac{1}{s-r}\sum_{t=r}^{s-1} \mathbf{x}_t\epsilon_t \right\|_2 \ge cJ_{min} \right) \\ \le & \sum_{\substack{1\le r<s\le n+1\\ s-r\ge n\delta_n}} P\left(\left\| \frac{1}{s-r}\sum_{t=r}^{s-1} \mathbf{x}_t\epsilon_t \right\|_1 \ge cJ_{min} \right). \end{aligned}$$

Define $Q_m = \frac{1}{s-r}\sum_{t=r}^{s-1} x_{t,m}\epsilon_i$ for $1\le m\le p$, so $Q_m \sim \mathcal{N}\left(0, \frac{\sigma^2}{(s-r)^2}\sum_{t=r}^{s-1} x_{t,m}^2\right)$.

$$\begin{aligned} & \sum_{\substack{1\le r<s\le n+1\\ s-r\ge n\delta_n}} P\left(\sum_{m=1}^{p} |Q_m| \ge cJ_{min} \right) \\ \le & \sum_{\substack{1\le r<s\le n+1\\ s-r\ge n\delta_n}} P\left(\max_{1\le m\le p} |Q_m| \ge \frac{cJ_{min}}{p} \right) \\ \le & \sum_{\substack{1\le r<s\le n+1\\ s-r\ge n\delta_n}} \sum_{m=1}^{p} P\left(|Q_m| \ge \frac{cJ_{min}}{p} \right) \\ \le & \sum_{\substack{1\le r<s\le n+1\\ s-r\ge n\delta_n}} \sum_{m=1}^{p} \exp\left(-\frac{(cJ_{min}/p)^2}{\frac{2\sigma^2}{(s-r)^2}\sum_{t=r}^{s-1} x_{t,m}^2} \right) \to 0, \end{aligned}$$

in which the last step holds as we have n^2p terms in the summation and A3.

So $\max_{\substack{1\le r<s\le n+1\\ s-r\ge n\delta_n}} \left\| \frac{1}{s-r}\sum_{t=s}^{r-1} \mathbf{x}_t\epsilon_t \right\|_2 = o_P(J_{min})$. □

Lemma 5 implies that, under A3, the interference of noise is negligible compared with J_{min} as $n \to \infty$.

A.2 Proof of Proposition 1

The proof follows closely from the proof of Theorem 3.1(i) in [40] and the proof of Proposition 3 in [39].

Since $P(\max_{1\leq k\leq K^*} |\hat{t}_k - t_k^*| > n\delta_n) \leq \sum_{k=1}^{K^*} P(|\hat{t}_k - t_k^*| > n\delta_n)$, it suffices to show $P\left(|\hat{t}_k - t_k^*| > n\delta_n\right) \to 0$ for each k. Define $A_{n,k} = \left\{|\hat{t}_k - t_k^*| \geq n\delta_n\right\}$, $C_n = \left\{\max_{0\leq k\leq K^*} |\hat{t}_k - t_k^*| < I_{min}/2\right\}$, it suffices to show that $P(A_{n,k} \cap C_n) \to 0$ and that $P(A_{n,k} \cap \bar{C}_n) \to 0$. This can be proved by repeatedly using Lemma 4 and Lemma 5.

A.2.1 Prove: $P(A_{n,k} \cap C_n) \to 0$.

Define $A_{n,k}^+ = \{t_k^* - \hat{t}_k \geq n\delta_n\}$, $A_{n,k}^- = \{\hat{t}_k - t_k^* \geq n\delta_n\}$, it suffices to prove $P(A_{n,k}^+ \cap C_n) \to 0$ and $P(A_{n,k}^- \cap C_n) \to 0$.

Let us first prove $P(A_{n,k}^+ \cap C_n) \to 0$. On the event $A_{n,k}^+ \cap C_n$, we have $t_{k-1}^* < \hat{t}_k < t_{k+1}^*$, for all $k \in \{1, 2, ..., K^*\}$.

Let $\tilde{\lambda}_n$ denote $\lambda_n \left(\gamma + (1-\gamma)\sqrt{p}\right)$. Applying (A.1), with $t = t_k^*$ and $t = \hat{t}_k$ respectively, gives $\frac{1}{n}\| \sum_{t=t_k^*}^{n} \mathbf{x}_t(y_t - \mathbf{x}_t^T\hat{\boldsymbol{\beta}}_t)\|_2 \leq \frac{\tilde{\lambda}_n}{2}$ and $\frac{1}{n}\| \sum_{t=\hat{t}_k}^{n} \mathbf{x}_t(y_t - \mathbf{x}_t^T\hat{\boldsymbol{\beta}}_t)\|_2 \leq \frac{\tilde{\lambda}_n}{2}$.

Using triangle inequality, we have

$$\begin{aligned}
\tilde{\lambda}_n &\geq \frac{1}{n}\left\|\sum_{t=\hat{t}_k}^{t_k^*-1} \mathbf{x}_t\left(y_t - \mathbf{x}_t^T\hat{\boldsymbol{\beta}}_t\right)\right\|_2 \\
&\geq \left\|\frac{1}{n}\sum_{t=\hat{t}_k}^{t_k^*-1} \mathbf{x}_t\mathbf{x}_t^T\left(\boldsymbol{\alpha}_{k+1}^* - \boldsymbol{\alpha}_k^*\right)\right\|_2 - \left\|\frac{1}{n}\sum_{t=\hat{t}_k}^{t_k^*-1} \mathbf{x}_t\mathbf{x}_t^T\left(\hat{\boldsymbol{\alpha}}_{k+1} - \boldsymbol{\alpha}_{k+1}^*\right)\right\|_2 - \left\|\frac{1}{n}\sum_{t=\hat{t}_k}^{t_k^*-1} \mathbf{x}_t\epsilon_t\right\|_2 \\
&\equiv R_{n,1} - R_{n,2} - R_{n,3}.
\end{aligned}$$

Then we apply a technique that we will use frequently in the later part of proof,

$$\begin{aligned}
& P(A_{n,k}^{+} \cap C_n) \\
= \;& P\left(A_{n,k}^{+} \cap C_n \cap \left\{\tilde{\lambda}_n \geq \frac{1}{3}R_{n,1}\right\}\right) + P\left(A_{n,k}^{+} \cap C_n \cap \left\{\tilde{\lambda}_n < \frac{1}{3}R_{n,1}\right\}\right) \\
\leq \;& P\left(A_{n,k}^{+} \cap C_n \cap \left\{\tilde{\lambda}_n \geq \frac{1}{3}R_{n,1}\right\}\right) + P\left(A_{n,k}^{+} \cap C_n \cap \left\{\frac{2}{3}R_{n,1} < R_{n,2} + R_{n,3}\right\}\right) \\
\leq \;& P\left(A_{n,k}^{+} \cap C_n \cap \left\{\tilde{\lambda}_n \geq \frac{1}{3}R_{n,1}\right\}\right) \\
& + P\left(A_{n,k}^{+} \cap C_n \cap \left\{\frac{2}{3}R_{n,1} < R_{n,2} + R_{n,3}\right\} \cap \left\{R_{n,2} \geq \frac{1}{3}R_{n,1}\right\}\right) \\
& + P\left(A_{n,k}^{+} \cap C_n \cap \left\{\frac{2}{3}R_{n,1} < R_{n,2} + R_{n,3}\right\} \cap \left\{R_{n,2} < \frac{1}{3}R_{n,1}\right\}\right) \\
\leq \;& P\left(A_{n,k}^{+} \cap C_n \cap \left\{\tilde{\lambda}_n \geq \frac{1}{3}R_{n,1}\right\}\right) + P\left(A_{n,k}^{+} \cap C_n \cap \left\{R_{n,2} \geq \frac{1}{3}R_{n,1}\right\}\right) \\
& + P\left(A_{n,k}^{+} \cap C_n \cap \left\{R_{n,3} \geq \frac{1}{3}R_{n,1}\right\}\right).
\end{aligned}$$

For the first term,

$$\begin{aligned}
& P\left(A_{n,k}^{+} \cap C_n \cap \left\{\tilde{\lambda}_n \geq \frac{1}{3}R_{n,1}\right\}\right) \\
\leq \;& P\left(A_{n,k}^{+} \cap \left\{\tilde{\lambda}_n \geq \frac{1}{3}R_{n,1}\right\}\right) \\
= \;& P\left(\left\{t_k^* - \hat{t}_k \geq n\delta_n\right\} \cap \left\{\frac{3n\tilde{\lambda}_n}{t_k^* - \hat{t}_k} \geq \left\|\frac{1}{t_k^* - \hat{t}_k}\sum_{t=\hat{t}_k}^{t_k^*-1} \mathbf{x}_t\mathbf{x}_t^T(\boldsymbol{\alpha}_{k+1}^* - \boldsymbol{\alpha}_k^*)\right\|_2\right\}\right) \\
\leq \;& P\left(\left\{t_k^* - \hat{t}_k \geq n\delta_n\right\} \cap \left\{\frac{3\tilde{\lambda}_n}{J_{min}\delta_n} \geq \underline{\mu}_{1n,k}\right\}\right),
\end{aligned}$$

where the last step is due to the fact that $\|AB\|_2 \geq \mu_{min}(A)\|B\|_2$ for symmetric positive semidefinite matrix A. And $\underline{\mu}_{1n,k} = \mu_{min}\left(\frac{1}{t_k^*-\hat{t}_k}\sum_{t=\hat{t}_k}^{t_k^*-1}\mathbf{x}_t\mathbf{x}_t^T\right)$. From Assumption 1 we know that $\underline{\mu}_{1n,k}$ is lower bounded by $l > 0$ as $n \to \infty$, and we choose λ_n such that $\frac{\lambda_n}{J_{min}\delta_n} \to 0$, as $n \to \infty$, we have that $P\left(A_{n,k}^{+} \cap C_n \cap \left\{\tilde{\lambda}_n \geq \frac{1}{3}R_{n,1}\right\}\right) \to 0$ as $n \to \infty$.

For the second term,

$$\begin{aligned}
& P\left(A_{n,k}^{+} \cap C_n \cap \left\{R_{n,2} \geq \frac{1}{3} R_{n,1}\right\}\right) \\
\leq \;& P\left(A_{n,k}^{+} \cap C_n \cap \left\{\bar{\mu}_{1n,k} \|\hat{\boldsymbol{\alpha}}_{k+1} - \boldsymbol{\alpha}_{k+1}^{*}\|_2 \geq \frac{1}{3} \underline{\mu}_{1n,k} \|\boldsymbol{\alpha}_{k+1}^{*} - \boldsymbol{\alpha}_{k}^{*}\|_2\right\}\right),
\end{aligned}$$

where $\bar{\mu}_{1n,k} = \mu_{max}\left(\frac{1}{t_k^* - \hat{t}_k} \sum_{t=\hat{t}_k}^{t_k^*-1} \mathbf{x}_t \mathbf{x}_t^T\right)$. Note that in the time interval $t \in [t_k^*, (t_k^* + t_{k+1}^*)/2 - 1]$, $\hat{\boldsymbol{\beta}}_t = \hat{\boldsymbol{\alpha}}_{k+1}$ and $\boldsymbol{\beta}_t^* = \boldsymbol{\alpha}_{k+1}^*$. Using Lemma 4, we have

$$\tilde{\lambda}_n \geq \left\| \frac{1}{n} \sum_{t=t_k^*}^{(t_k^*+t_{k+1}^*)/2-1} \mathbf{x}_t \mathbf{x}_t^T (\hat{\boldsymbol{\alpha}}_{k+1} - \boldsymbol{\alpha}_{k+1}^*) \right\|_2 - \left\| \frac{1}{n} \sum_{t=t_k^*}^{(t_k^*+t_{k+1}^*)/2-1} \mathbf{x}_t \epsilon_t \right\|_2 .$$

Since $t_{k+1}^* - t_k^* \geq I_{min}$, so we have

$$\begin{aligned}
& \frac{n\tilde{\lambda}_n}{(t_{k+1}^* - t_k^*)/2} + \left\| \frac{2}{t_{k+1}^* - t_k^*} \sum_{t=t_k^*}^{(t_k^*+t_{k+1}^*)/2-1} \mathbf{x}_t \epsilon_t \right\|_2 \\
\geq \;& \left\| \frac{2}{t_{k+1}^* - t_k^*} \sum_{t=t_k^*}^{(t_k^*+t_{k+1}^*)/2-1} \mathbf{x}_t \mathbf{x}_t^T (\hat{\boldsymbol{\alpha}}_{k+1} - \boldsymbol{\alpha}_{k+1}^*) \right\|_2 \\
\geq \;& \underline{\mu}_{2n,k} \|\hat{\boldsymbol{\alpha}}_{k+1} - \boldsymbol{\alpha}_{k+1}^*\|_2,
\end{aligned}$$

so $\|\hat{\boldsymbol{\alpha}}_{k+1} - \boldsymbol{\alpha}_{k+1}^*\|_2 \leq \underline{\mu}_{2n,k}^{-1} \left[\frac{2n\tilde{\lambda}_n}{I_{min}} + \left\| \frac{2}{t_{k+1}^* - t_k^*} \sum_{t=t_k^*}^{(t_k^*+t_{k+1}^*)/2-1} \mathbf{x}_t \epsilon_t \right\|_2\right]$. Then we can bound the second term by

$$\begin{aligned}
& P\left(C_n \cap \left\{\bar{\mu}_{1n,k} \|\hat{\boldsymbol{\alpha}}_{k+1} - \boldsymbol{\alpha}_{k+1}^*\|_2 \geq \frac{1}{3} \underline{\mu}_{1n,k} \|\boldsymbol{\alpha}_{k+1}^* - \boldsymbol{\alpha}_k^*\|_2\right\}\right) \\
\leq \;& P\left(\frac{2n\tilde{\lambda}_n}{I_{min}} \geq \bar{\mu}_{1n,k}^{-1} \underline{\mu}_{1n,k} \underline{\mu}_{2n,k} \left\|\boldsymbol{\alpha}_{k+1}^* - \boldsymbol{\alpha}_k^*\right\|_2 / 6\right) \\
& + P\left(\left\| \frac{2}{t_{k+1}^* - t_k^*} \sum_{t=t_k^*}^{(t_k^*+t_{k+1}^*)/2-1} \mathbf{x}_t \epsilon_t \right\|_2 \geq \bar{\mu}_{1n,k}^{-1} \underline{\mu}_{1n,k} \underline{\mu}_{2n,k} \left\|\boldsymbol{\alpha}_{k+1}^* - \boldsymbol{\alpha}_k^*\right\|_2 / 6\right).
\end{aligned}$$

From A2 and the way we choose λ_n, $\frac{n\lambda_n}{I_{min}J_{min}} \to 0$. From Lemma 5, and $I_{min}/(n\delta_n) \to \infty$, we have $\left\|\frac{2}{t^*_{k+1}-t^*_k}\sum_{t=t^*_k}^{(t^*_k+t^*_{k+1})/2-1}\mathbf{x}_t\epsilon_t\right\|_2 = o_P(J_{min})$. So $P\left(A^+_{n,k}\cap C_n\cap\left\{R_{n,2}\geq\frac{1}{3}R_{n,1}\right\}\right)\to 0$.

For the third term,

$$\begin{aligned}
& P\left(A^+_{n,k}\cap C_n\cap\left\{R_{n,3}\geq\frac{1}{3}R_{n,1}\right\}\right) \\
\leq\ & P\left(A^+_{n,k}\cap\left\{\left\|\frac{1}{t^*_k-\hat{t}_k}\sum_{t=\hat{t}_k}^{t^*_k-1}\mathbf{x}_t\epsilon_t\right\|_2\geq\frac{1}{3}\left\|\frac{1}{t^*_k-\hat{t}_k}\sum_{t=\hat{t}_k}^{t^*_k-1}\mathbf{x}_t\mathbf{x}_t^T(\boldsymbol{\alpha}^*_{k+1}-\boldsymbol{\alpha}^*_k)\right\|_2\right\}\right) \\
\leq\ & P\left(A^+_{n,k}\cap\left\{\left\|\frac{1}{t^*_k-\hat{t}_k}\sum_{t=\hat{t}_k}^{t^*_k-1}\mathbf{x}_t\epsilon_t\right\|_2\geq\frac{1}{3}\underline{\mu}_{1n,k}J_{min}\right\}\right) \\
\to\ & 0,
\end{aligned}$$

since $\underline{\mu}_{1n,k}J_{min}\geq l>0$ and $\|\frac{1}{t^*_k-\hat{t}_k}\sum_{t=\hat{t}_k}^{t^*_k-1}\mathbf{x}_t\epsilon_t\|_2 = o_P(J_{min})$ in $A^+_{n,k}$.

The proof for $P(A^-_{n,k}\cap C_n)\to 0$ follows from similar steps. Thus we prove that $P(A_{n,k}\cap C_n)\to 0$ as n goes to infinity.

A.2.2 Prove: $P(A_{n,k}\cap\bar{C}_n)\to 0$.

It suffices to prove that $P(A^+_{n,k}\cap\bar{C}_n)\to 0$ and $P(A^-_{n,k}\cap\bar{C}_n)\to 0$. We first prove $P(A^+_{n,k}\cap\bar{C}_n)\to 0$ and the other case can be proved similarly.

Define

$$\begin{aligned}
D_n^{(l)} &= \{\exists k\in\{1,...,K^*\},\hat{t}_k\leq t^*_{k-1}\}\cap\bar{C}_n, \\
D_n^{(m)} &= \{\forall k\in\{1,...,K^*\},t^*_{k-1}<\hat{t}_k<t^*_{k+1}\}\cap\bar{C}_n, \\
D_n^{(l)} &= \{\exists k\in\{1,...,K^*\},\hat{t}_k\geq t^*_{k+1}\}\cap\bar{C}_n,
\end{aligned}$$

then $P(A^+_{n,k}\cap\bar{C}_n) = P(A^+_{n,k}\cap D_n^{(l)}) + P(A^+_{n,k}\cap D_n^{(m)}) + P(A^+_{n,k}\cap D_n^{(r)})$.

Here we only consider the case $P(A_{n,k}^{+} \cap D_n^{(m)})$, and the proof for other two cases follows similarly.

$$\begin{aligned}
& P(A_{n,k}^{+} \cap D_n^{(m)}) \\
= \; & P(A_{n,k}^{+} \cap \{\hat{t}_{k+1} - t_k^* \geq I_{min}/2\} \cap D_n^{(m)}) \\
& + P(A_{n,k}^{+} \cap \{\hat{t}_{k+1} - t_k^* < I_{min}/2\} \cap D_n^{(m)}) \\
\leq \; & P(A_{n,k}^{+} \cap \{\hat{t}_{k+1} - t_k^* \geq I_{min}/2\} \cap D_n^{(m)}) \\
& + P(A_{n,k}^{+} \cap \{t_{k+1}^* - \hat{t}_{k+1} \geq I_{min}/2\} \cap D_n^{(m)}),
\end{aligned}$$

where the last step is from $t_{k+1}^* - \hat{t}_{k+1} = (t_{k+1}^* - t_k^*) - (\hat{t}_{k+1} - t_k^*) \geq I_{min} - I_{min}/2 = I_{min}/2$. Notice that

$$\begin{aligned}
& \left\{A_{n,k}^{+} \cap \left\{t_{k+1}^* - \hat{t}_{k+1} \geq I_{min}/2\right\} \cap D_n^{(m)}\right\} \\
& \subset \cup_{i=k+1}^{K^*-1} \left(\left\{t_i^* - \hat{t}_i \geq I_{min}/2\right\} \cap \left\{\hat{t}_{i+1} - t_i^* \geq I_{min}/2\right\} \cap D_n^{(m)}\right),
\end{aligned}$$

we have

$$\begin{aligned}
& P(A_{n,k}^{+} \cap D_n^{(m)}) \\
\leq \; & P(A_{n,k}^{+} \cap \{\hat{t}_{k+1} - t_k^* \geq I_{min}/2\} \cap D_n^{(m)}) \\
& + \sum_{i=k+1}^{K^*-1} P\left(\left\{t_i^* - \hat{t}_i \geq I_{min}/2\right\} \cap \left\{\hat{t}_{i+1} - t_i^* \geq I_{min}/2\right\} \cap D_n^{(m)}\right). \qquad \text{(A.7)}
\end{aligned}$$

To bound the first term in (A.7), we apply Lemma 4 with $t = \hat{t}_k$ and $t = t_k^*$ respectively, so $\frac{1}{n}\left\|\sum_{i=\hat{t}_k}^{n} \mathbf{x_i}(y_i - \mathbf{x_i}^T \hat{\boldsymbol{\beta}}_{\boldsymbol{i}})\right\|_2 \leq \frac{\tilde{\lambda}_n}{2}$ and $\frac{1}{n}\left\|\sum_{i=t_k^*}^{n} \mathbf{x_i}(y_i - \mathbf{x_i}^T \hat{\boldsymbol{\beta}}_{\boldsymbol{i}})\right\|_2 \leq \frac{\tilde{\lambda}_n}{2}$. Using triangle inequality, we have $\frac{\tilde{\lambda}_n}{t_k^* - \hat{t}_k} \geq \frac{1}{t_k^* - \hat{t}_k}\| - \sum_{i=\hat{t}_k}^{t_k^*-1} \mathbf{x_i}\mathbf{x_i}^T(\hat{\boldsymbol{\alpha}}_{\boldsymbol{k+1}} - \boldsymbol{\alpha}_{\boldsymbol{k}}^{\boldsymbol{*}}) + \sum_{i=\hat{t}_k}^{t_k^*-1} \mathbf{x_i}\epsilon_i\|_2 \geq$

$\underline{\mu}_{1n,k}\|\boldsymbol{\hat{\alpha}_{k+1}} - \boldsymbol{\alpha^*_k}\|_2 - \|\frac{1}{t^*_k - \hat{t}_k}\sum_{i=\hat{t}_k}^{t^*_k-1}\mathbf{x_i}\epsilon_i\|_2$. So we have

$$\begin{aligned} & \|\boldsymbol{\hat{\alpha}_{k+1}} - \boldsymbol{\alpha^*_k}\|_2 \\ \leq \;& \underline{\mu}^{-1}_{1n,k}\left(\frac{n\tilde{\lambda}_n}{t^*_k - \hat{t}_k} + \left\|\frac{1}{t^*_k - \hat{t}_k}\sum_{i=\hat{t}_k}^{t^*_k-1}\mathbf{x_i}\epsilon_i\right\|_2\right). \end{aligned} \tag{A.8}$$

Following the same procedure, we apply Lemma 4 with $t = \hat{t}_{k+1}$ and $t = t^*_k$ respectively, we have

$$\begin{aligned} & \|\boldsymbol{\hat{\alpha}_{k+1}} - \boldsymbol{\alpha^*_{k+1}}\|_2 \\ \leq \;& \underline{\mu}^{-1}_{3n,k}\left(\frac{n\tilde{\lambda}_n}{\hat{t}_{k+1} - t^*_k} + \left\|\frac{1}{\hat{t}_{k+1} - t^*_k}\sum_{i=t^*_k}^{\hat{t}_{k+1}-1}\mathbf{x_i}\epsilon_i\right\|_2\right), \end{aligned} \tag{A.9}$$

where $\underline{\mu}^{-1}_{3n,k} = \mu_{min}\left(\frac{1}{\hat{t}_{k+1} - t^*_k}\sum_{i=t^*_k}^{\hat{t}_{k+1}-1}\mathbf{x_i}\mathbf{x_i}^T\right)$. After combining (A.8) and (A.9), we conclude that the event

$$\begin{aligned} E_{n,k} \;=\; & \big\{\|\boldsymbol{\alpha^*_{k+1}} - \boldsymbol{\alpha^*_k}\|_2 \\ & \leq n\tilde{\lambda}_n\left(\underline{\mu}^{-1}_{1n,k}\frac{1}{t^*_k - \hat{t}_k} + \underline{\mu}^{-1}_{3n,k}\frac{1}{\hat{t}_{k+1} - t^*_k}\right) \\ & + \underline{\mu}^{-1}_{1n,k}\left\|\frac{1}{t^*_k - \hat{t}_k}\sum_{i=\hat{t}_k}^{t^*_k-1}\mathbf{x_i}\epsilon_i\right\|_2 \\ & + \underline{\mu}^{-1}_{3n,k}\left\|\frac{1}{\hat{t}_{k+1} - t^*_k}\sum_{i=t^*_k}^{\hat{t}_{k+1}-1}\mathbf{x_i}\epsilon_i\right\|_2\Bigg\} \end{aligned}$$

occurs with probability one. So for the first term in (A.7), using the same technique as above, we have

$$
\begin{aligned}
& P\left(A^{+}_{n,k} \cap \left\{\hat{t}_{k+1} - t^*_k \geq I_{min}/2\right\} \cap D_n^{(m)}\right) \\
= \;& P\left(E_{n,k} \cap A^{+}_{n,k} \cap \{\hat{t}_{k+1} - t^*_k \geq I_{min}/2\} \cap D_n^{(m)}\right) \\
\leq \;& P\left(E_{n,k} \cap \left\{t^*_k - \hat{t}_k \geq n\delta_n\right\} \cap \left\{\hat{t}_{k+1} - t^*_k \geq I_{min}/2\right\}\right) \\
\leq \;& P\left((\gamma + (1-\gamma)\sqrt{p})\left(n\lambda_n \delta_n^{-1} \underline{\mu}^{-1}_{1n,k} + \frac{2n\lambda_n}{I_{min}} \underline{\mu}^{-1}_{3n,k}\right) \geq \left\|\boldsymbol{\alpha}^*_{\boldsymbol{k+1}} - \boldsymbol{\alpha}^*_{\boldsymbol{k}}\right\|_2 /3\right) \\
& + P\left(\left\{\underline{\mu}^{-1}_{1n,k}\left\|\frac{1}{t^*_k - \hat{t}_k}\sum_{i=\hat{t}_k}^{t^*_k - 1} \mathbf{x_i}\epsilon_i\right\|_2 \geq \left\|\boldsymbol{\alpha}^*_{\boldsymbol{k+1}} - \boldsymbol{\alpha}^*_{\boldsymbol{k}}\right\|_2 /3\right\} \cap \left\{t^*_k - \hat{t}_k \geq n\delta_n\right\}\right) \\
& + P\left(\left\{\underline{\mu}^{-1}_{3n,k}\left\|\frac{1}{\hat{t}_{k+1} - t^*_k}\sum_{i=t^*_k}^{\hat{t}_{k+1} - 1} \mathbf{x_i}\epsilon_i\right\|_2 \geq \left\|\boldsymbol{\alpha}^*_{\boldsymbol{k+1}} - \boldsymbol{\alpha}^*_{\boldsymbol{k}}\right\|_2 /3\right\} \cap \left\{\hat{t}_{k+1} - t^*_k \geq I_{min}/2\right\}\right) \\
& \to 0.
\end{aligned}
$$

In the the last step above, the first term tends to zero since $\frac{\lambda_n}{J_{min}\delta_n} \to 0$ as $n \to \infty$ and $\frac{n\lambda_n}{I_{min}J_{min}} \to 0$ from A1 and A3, the second tends to zero since $\left\|\frac{1}{t^*_k - \hat{t}_k}\sum_{i=\hat{t}_k}^{t^*_k-1} \mathbf{x_i}\epsilon_i\right\|_2 = o_P(J_{min})$ from Lemma 4, the third term tends to zero since $\frac{I_{min}}{n\delta_n} \to \infty$ and hence $\left\|\frac{1}{\hat{t}_{k+1} - t^*_k}\sum_{i=t^*_k}^{\hat{t}_{k+1}-1} \mathbf{x_i}\epsilon_i\right\|_2 = o_P(J_{min})$ from Lemma 5.

Using similar procedure, we can show that other terms in (A.7) go to zero.

Combining results in Appendix A.2.1 and Appendix A.2.2, we complete the proof.

A.3 Proof of Proposition 2

We prove this proposition by dividing the problem into four cases: 1) $\hat{t}_{k-1} < t^*_{k-1}$ and $\hat{t}_k < t^*_k$, 2) $t^*_{k-1} \leq \hat{t}_{k-1}$ and $\hat{t}_k < t^*_k$, 3)$\hat{t}_{k-1} < t^*_{k-1}$ and $t^*_k \leq \hat{t}_k$, 4) $t^*_{k-1} \leq \hat{t}_{k-1}$ and $t^*_k \leq \hat{t}_k$. We then prove each case by using Lemma 1 repeatedly. Here we only give the proof for case 1) and other cases follow similarly.

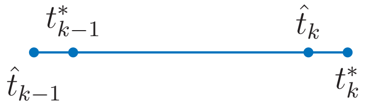

Figure A.1: Case 1): $\hat{t}_{k-1} < t^*_{k-1}$ and $\hat{t}_k < t^*_k$

Using Lemma 4 by setting $t = t^*_{k-1}$ and $t = \hat{t}_k$ respectively, we have

$$\begin{aligned}\tilde{\lambda}_n &\geq \left\| \frac{1}{n} \sum_{t=t^*_{k-1}}^{\hat{t}_k - 1} \mathbf{x}_t (y_t - \mathbf{x}_t^T \hat{\boldsymbol{\alpha}}_k) \right\|_2 \\ &\geq \frac{\hat{t}_k - t^*_{k-1}}{n} \left\| \frac{1}{\hat{t}_k - t^*_{k-1}} \sum_{t=t^*_{k-1}}^{\hat{t}_k - 1} \mathbf{x}_t \mathbf{x}_t^T (\boldsymbol{\alpha}^*_k - \hat{\boldsymbol{\alpha}}_k) \right\|_2 \\ &\quad - \frac{\hat{t}_k - t^*_{k-1}}{n} \left\| \frac{1}{\hat{t}_k - t^*_{k-1}} \sum_{t=t^*_{k-1}}^{\hat{t}_k - 1} \mathbf{x}_t \epsilon_t \right\|_2 .\end{aligned}$$

By Proposition 1, we have $\hat{t}_k - t^*_{k-1} \geq (t^*_k - t^*_{k-1}) - n\delta_n \geq I_{min} - n\delta_n$, and by A2, $\hat{t}_k - t^*_{k-1} > n\delta_n$ as $n \to \infty$. Then using A1, Lemma 5, and the fact that $I_k - n\delta_n \leq \hat{t}_k - t^*_{k-1} \leq I_k$, where I_k is defined as $t^*_k - t^*_{k-1}$ and $\|AB\|_2 \geq \mu_{min}(A)\|B\|_2$ for symmetric positive semidefinite matrix A, we have

$$\begin{aligned}\tilde{\lambda}_n &\geq \frac{\hat{t}_k - t^*_{k-1}}{n} l \left\| \boldsymbol{\alpha}^*_k - \hat{\boldsymbol{\alpha}}_k \right\|_2 - \frac{\hat{t}_k - t^*_{k-1}}{n} o_P(J_{min}) \\ &\geq \frac{I_k - n\delta_n}{n} l \left\| \boldsymbol{\alpha}^*_k - \hat{\boldsymbol{\alpha}}_k \right\|_2 - \frac{I_k}{n} o_P(J_{min}).\end{aligned}$$

So we have

$$\|\boldsymbol{\alpha}^*_k - \hat{\boldsymbol{\alpha}}_k\|_2 \leq \frac{n\tilde{\lambda}_n + I_k o_P(J_{min})}{(I_k - n\delta_n) l} \leq \frac{n\tilde{\lambda}_n + I_{min} o_P(J_{min})}{(I_{min} - n\delta_n) l},$$

in which the last step is due to the fact that $I_{min} \leq I_k$.

After studying for all four cases, we compare and pick the largest one, and thus complete our proof.

A.4 Proof of Proposition 3

The proof follows closely from the proof of Theorem 3.2 in [40] and the proof of Proposition 4 in [39]. Due to space limitation, we provide only an outline of the proof.

It suffices to show that $P(\varepsilon(\hat{T}_{\hat{K}}||T^*) > n\delta_n\} \cap \{K^* < \hat{K} \leq K_{max}\}) \to 0$ as $n \to \infty$.

$$\begin{aligned}
& P(\{\varepsilon(\hat{T}_{\hat{K}}||T^*) > n\delta_n\} \cap \{K^* < \hat{K} \leq K_{max}\}) \\
\leq & \sum_{m=K^*+1}^{K_{max}} P(\varepsilon(\hat{T}_m||T^*) > n\delta_n) \\
\leq & \sum_{m=K^*+1}^{K_{max}} \sum_{k=1}^{K^*} P(\forall l, 1 \leq l \leq m, |\hat{t}_l - t_k^*| > n\delta_n) \\
= & \sum_{m=K^*+1}^{K_{max}} \sum_{k=1}^{K^*} [P(F_{m,k,1}) + P(F_{m,k,2}) + P(F_{m,k,3})],
\end{aligned}$$

where $F_{m,k,1} = \{\forall l, 1 \leq l \leq m, |\hat{t}_l - t_k^*| > n\delta_n, \hat{t}_l < t_k^*\}$, $F_{m,k,2} = \{\forall l, 1 \leq l \leq m, |\hat{t}_l - t_k^*| > n\delta_n, \hat{t}_l > t_k^*\}$, $F_{m,k,3} = \{\exists l, 1 \leq l \leq m, |\hat{t}_l - t_k^*| > n\delta_n, |\hat{t}_{l+1} - t_k^*| > n\delta_n, \hat{t}_l < t_k^* < \hat{t}_{l+1}\}$.

It suffices to prove $P(F_{m,k,1}) \to 0$, $P(F_{m,k,2}) \to 0$ and $P(F_{m,k,3}) \to 0$. This can be proved by repeatedly using Lemma 4 and Lemma 5.

A.5 Proof of Proposition 4

The proof follows closely from the proof of Theorem 3.3 in [40] and the proof of Proposition 4 in [39].

We can prove this proposition by contradiction. Let $\check{K}$ be the estimated number of points and $\check{K} < K^*$. Let $\check{T}_{\check{K}} = \{\check{t}_1, \check{t}_2, \cdots, \check{t}_{\check{K}}\}$ be the set of all $\check{K}$ locations of estimated change-points. Let $\check{\boldsymbol{\alpha}} = \{\check{\boldsymbol{\alpha}}_1, \check{\boldsymbol{\alpha}}_1, \cdots, \check{\boldsymbol{\alpha}}_{\check{K}}\}$ be the estimated coefficients. Let $\check{\boldsymbol{\beta}}_t$ be the estimated coefficients at time t. $\check{K}$, $\check{T}_{\check{K}}$ and $\check{\boldsymbol{\alpha}}$ achieve the minimum of Problem (2.7).

It suffices to prove that there exists some other set of change-points $\hat{T}_{\hat{K}} = \{\hat{t}_1, \hat{t}_2, \cdots, \hat{t}_{\hat{K}}\}$ and coefficients $\hat{\boldsymbol{\alpha}} = \{\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_1, \cdots, \hat{\boldsymbol{\alpha}}_{\hat{K}}\}$ that can achieve a smaller value. Let $\hat{\boldsymbol{\beta}}_t$ be the corresponding coefficients at time t.

Define $\Delta(\check{T}_{\check{K}}; \hat{T}_{\hat{K}}) = \frac{1}{n} \sum_{j=1}^{\check{K}+1} \sum_{t=\check{t}_{j-1}}^{\check{t}_j - 1} \left(y_t - \check{\boldsymbol{\alpha}}_j{}^T \mathbf{x}_t\right)^2 - \frac{1}{n} \sum_{j=1}^{\hat{K}+1} \sum_{t=\hat{t}_{j-1}}^{\hat{t}_j - 1} \left(y_t - \hat{\boldsymbol{\alpha}}_j{}^T \mathbf{x}_t\right)^2 + \lambda_n(\gamma \sum_{j=1}^{\check{K}} \|\check{\boldsymbol{\alpha}}_{j+1} - \check{\boldsymbol{\alpha}}_j\|_2 - \gamma \sum_{j=1}^{\hat{K}} \|\hat{\boldsymbol{\alpha}}_{j+1} - \hat{\boldsymbol{\alpha}}_j\|_2 + (1-\gamma) \sum_{j=1}^{\check{K}} \|\check{\boldsymbol{\alpha}}_{j+1} - \check{\boldsymbol{\alpha}}_j\|_1 - (1-\gamma) \sum_{j=1}^{\hat{K}} \|\hat{\boldsymbol{\alpha}}_{j+1} - \hat{\boldsymbol{\alpha}}_j\|_1)$.

From simple analysis, we know that if $\check{K} < K^*$, then there exist some t_k^* such that there is no estimated change-points falling into the neighborhood of range $I_{min}/2$. That is $\varepsilon(\check{T}_{\check{K}} || \{t_k^*\}) \geq I_{min}/2$ for some k.

Then we prove that $\check{K}$ such that $\check{K} < K^*$ cannot be an optimal solution. We prove this by dividing the problem into four cases: 1) $\check{t}_{l-1} \leq t_{k-1}^*$ and $\check{t}_l \geq t_{k+1}^*$, 2) $\check{t}_{l-1} > t_{k-1}^*$ and $\check{t}_l < t_{k+1}^*$, 3) $\check{t}_{l-1} \leq t_{k-1}^*$ and $\check{t}_l < t_{k+1}^*$, 4) $\check{t}_{l-1} > t_{k-1}^*$ and $\check{t}_l \geq t_{k+1}^*$, where $\check{t}_{l-1}$ and $\check{t}_l$ are the change-points in $\check{T}_{\check{K}}$ nearest to t_k^* on left side and right side respectively.

Here, we prove only case 1), as other cases are similar.

There exist $\hat{t}_t = \check{t}_t$, for $t \leq l-1$, $\hat{t}_{l+1} = t_k^*$, $\hat{t}_{l+2} = t_{k+1}^*$, and $\hat{t}_t = \check{t}_{t-3}$, for $t \geq l+3$. Furthermore, there exist $\hat{\boldsymbol{\alpha}}_t = \check{\boldsymbol{\alpha}}_t$, for $t \leq l$, $\hat{\boldsymbol{\alpha}}_{l+1} = \boldsymbol{\alpha}_k^*$, $\hat{\boldsymbol{\alpha}}_{l+2} = \boldsymbol{\alpha}_{k+1}^*$, and $\hat{\boldsymbol{\alpha}}_t = \check{\boldsymbol{\alpha}}_{t-3}$, for $t \geq l+3$.

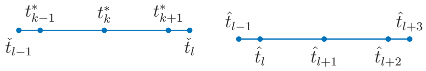

Figure A.2: Illustration of case 1)

So we have $\Delta(\check{T}_{\check{K}}; \hat{T}_{\hat{K}}) \geq \frac{1}{n} \sum_{t=t_{k-1}^*}^{t_{k+1}^*-1} [(y_t - \check{\boldsymbol{\alpha}}_{l-1}^T \mathbf{x}_t)^2 - \epsilon_t^2] - \tilde{\lambda}_n (\sum_{j=l}^{l+2} \|\hat{\boldsymbol{\alpha}}_{j+1} - \hat{\boldsymbol{\alpha}}_j\|_2)$. Define $A = \frac{1}{n} \sum_{t=t_{k-1}^*}^{t_{k+1}^*-1} \left[(y_t - \check{\boldsymbol{\alpha}}_{l-1}^T \mathbf{x}_t)^2 - \epsilon_t^2\right]$. Define $I_i = t_i^* - t_{i-1}^*$, $I_{all} = t_{k+1}^* -$

t^*_{k-1}, $d^j_i = \boldsymbol{\alpha}^*_i - \boldsymbol{\alpha}^*_j$, $Q = \frac{1}{t^*_{k+1}-t^*_{k-1}} \sum_{t=t^*_{k-1}}^{t^*_{k+1}-1} \mathbf{x}_t\mathbf{x}_t^T$, $Q_i = \frac{1}{t^*_i-t^*_{i-1}} \sum_{t=t^*_{i-1}}^{t^*_i-1} \mathbf{x}_t\mathbf{x}_t^T$, $S = \frac{1}{t^*_{k+1}-t^*_{k-1}} \sum_{t=t^*_{k-1}}^{t^*_{k+1}-1} \mathbf{x}_t\epsilon_t$ and $S_i = \frac{1}{t^*_i-t^*_{i-1}} \sum_{t=t^*_{i-1}}^{t^*_i-1} \mathbf{x}_t\epsilon_t$. Notice that Q and Q_i are asymptotically nonsingular, and are positive definite.

Notice

$$\begin{aligned}\sum_{j=l}^{l+2} \|\hat{\boldsymbol{\alpha}}_{j+1} - \hat{\boldsymbol{\alpha}}_j\|_2 = \|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}^*_k\|_2 \\ +\|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}^*_{k+1}\|_2 + \|\boldsymbol{\alpha}^*_{k+1} - \boldsymbol{\alpha}^*_k\|_2.\end{aligned} \tag{A.10}$$

Now, we consider different cases of (A.10).

1) $\|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}^*_k\|_2 < \infty$ and $\|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}^*_{k+1}\|_2 < \infty$ as $n \to \infty$. Since $\tilde{\lambda}_n$, we have $\Delta(\check{T}_{\check{K}}; \hat{T}_{\hat{K}}) \geq A + o_P(1)$. Then our goal becomes to find the minimum value of A. The value of $\check{\boldsymbol{\alpha}}_l$ that can achieve minimum value of A is

$$\begin{aligned}\check{\boldsymbol{\alpha}}_l &= \left(\sum_{t=t^*_{k-1}}^{t^*_{k+1}-1} \mathbf{x}_t\mathbf{x}_t^T\right)^{-1} \sum_{t=t^*_{k-1}}^{t^*_{k+1}-1} \mathbf{x}_t y_t \\ &= Q^{-1}S + Q^{-1} \sum_{i=k-1}^{k} \frac{I_{i+1}}{I_{all}} Q_{i+1}\alpha^*_{i+1},\end{aligned}$$

and $\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}^*_{j+1} = Q^{-1}S + Q^{-1} \sum_{i=k-1}^{k} \frac{I_{i+1}}{I_{all}} Q_{i+1} d^{j+1}_{i+1}$. Then we have

$$\begin{aligned}A &= \frac{1}{n} \sum_{j=k-1}^{k} \left[(\boldsymbol{\alpha}^*_{j+1} - \check{\boldsymbol{\alpha}}_l)^T \left(\sum_{t=t^*_j}^{t^*_{j+1}-1} \mathbf{x}_t\mathbf{x}_t^T\right) (\boldsymbol{\alpha}^*_{j+1} - \check{\boldsymbol{\alpha}}_l) \right. \\ &\quad \left. +2(\boldsymbol{\alpha}^*_{j+1} - \check{\boldsymbol{\alpha}}_l)^T \sum_{t=t^*_j}^{t^*_{j+1}-1} \mathbf{x}_t\epsilon_t \right]. \\ &\equiv \frac{1}{n} \sum_{j=k-1}^{k} (A_{j,1} + A_{j,2}).\end{aligned}$$

For $j = k-1$, we have

$$\begin{aligned}\frac{1}{n}A_{k-1,1} = &\frac{I_k}{n}\left(S^TQ^{-1}Q_kQ^{-1}S\right)\\ &+\frac{2I_k}{n}\left(\frac{I_{k+1}}{I_{all}}{d^k_{k+1}}^TQ_{k+1}Q^{-1}Q_kQ^{-1}S\right)\\ &+\frac{I_k}{I_{all}^2}\left(\frac{I_{k+1}^2}{n}{d^k_{k+1}}^TQ_{k+1}Q^{-1}Q_kQ^{-1}Q_{k+1}d^k_{k+1}\right).\end{aligned}$$

Notice that Q and Q_i have only positive eigenvalues, and the eigenvalues are bounded. So we have $\frac{n}{I_{min}J_{min}^2}\left|\frac{I_k}{n}\left(S^TQ^{-1}Q_kQ^{-1}S\right)\right| = |o_P(1)|$, where c_1 is a positive constant, and $\frac{n}{I_{min}J_{min}^2}\left|\frac{2I_k}{n}\left(\frac{I_{k+1}}{I_{all}}{d^k_{k+1}}^TQ_{k+1}Q^{-1}Q_kQ^{-1}S\right)\right| = |o_P(1)|$, where c_2 is a positive constant.

So we have $\frac{1}{n}A_{k-1,1} = \frac{I_{min}J_{min}^2}{n}o_P(1) + \frac{I_k}{I_{all}^2}(\frac{I_{k+1}^2}{n}{d^k_{k+1}}^TQ_{k+1}Q^{-1}Q_kQ^{-1}Q_{k+1}d^k_{k+1})$ and $\frac{1}{n}A_{k,1} = \frac{I_{min}J_{min}^2}{n}o_P(1) + \frac{I_{k+1}}{I_{all}^2}(\frac{I_k^2}{n}{d^{k+1}_k}^TQ_kQ^{-1}Q_{k+1}Q^{-1}Q_kd^{k+1}_k)$.

Notice that $\frac{I_k}{I_{all}^2}(\frac{I_{k+1}^2}{n}{d^k_{k+1}}^TQ_{k+1}Q^{-1}Q_kQ^{-1}Q_{k+1}d^k_{k+1})+\frac{I_{k+1}}{I_{all}^2}(\frac{I_k^2}{n}{d^{k+1}_k}^TQ_kQ^{-1}Q_{k+1}Q^{-1}Q_kd^{k+1}_k) \geq c_3\frac{I_{min}J_{min}^2}{n}$, where c_3 is some positive constant. Similarly, we can show $\frac{1}{n}A_{k-1,2} = \frac{I_{min}J_{min}^2}{n}o_P(1)$ and $\frac{1}{n}A_{k,2} = \frac{I_{min}J_{min}^2}{n}o_P(1)$. As the result, we have

$$A = \frac{I_{min}J_{min}^2}{n}\left(c_3 + o_P(1)\right),$$

and c_3 is some positive constant. So $\hat{K}$, $\hat{T}_{\hat{K}}$ and $\hat{\alpha}$ can achieve a smaller value, which means $\check{K}$ cannot be an optimal solution.

2) $\|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}^*_k\|_2 \to \infty$ or $\|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}^*_{k+1}\|_2 \to \infty$ as $n \to \infty$.

If $\|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}^*_k\|_2 \to \infty$ as $n \to \infty$, since $J_{max} = O(1)$, then

$$\|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}^*_k\|_2 - J_{max} \leq \|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}^*_{k+1}\|_2,$$

$\|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}^*_{k+1}\|_2 \to \infty$. We also prove if $\|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}^*_{k+1}\|_2 \to \infty$, then $\|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}^*_k\|_2 \to \infty$.

If $\|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_k^*\|_2 \to \infty$ and $\|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_{k+1}^*\|_2 \to \infty$, then

$$\begin{aligned}\Delta(\check{T}_{\check{K}}; \hat{T}_{\hat{K}}) \geq & \frac{I_k}{n}(\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_k^*)Q_k(\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_k^*)^T \\ & + \frac{I_{k+1}}{n}(\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_{k+1}^*)Q_{k+1}(\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_{k+1}^*)^T \\ & - \tilde{\lambda}_n \|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_k^*\|_2 - \tilde{\lambda}_n \|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_{k+1}^*\|_2 \\ & - \tilde{\lambda}_n \|\boldsymbol{\alpha}_{k+1}^* - \boldsymbol{\alpha}_k^*\|_2.\end{aligned}$$

Since $\Theta(n) = I_{min} \leq I_k \leq n$, so $\frac{I_k}{n}$ and $\frac{I_{k+1}}{n}$ are bounded below. So we have

$$\begin{aligned}\Delta(\check{T}_{\check{K}}; \hat{T}_{\hat{K}}) \geq & c_4 \|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_k^*\|_2^2 + c_5 \|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_{k+1}^*\|_2^2 \\ & - \tilde{\lambda}_n \|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_k^*\|_2 - \tilde{\lambda}_n \|\check{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_{k+1}^*\|_2 \\ & - \tilde{\lambda}_n \|\boldsymbol{\alpha}_{k+1}^* - \boldsymbol{\alpha}_k^*\|_2,\end{aligned}$$

where c_4 and c_5 are some positive constants. So $\Delta(\check{T}_{\check{K}}; \hat{T}_{\hat{K}})$ is dominated by the quadratic term, hence $\Delta(\check{T}_{\check{K}}; \hat{T}_{\hat{K}}) \to \infty$, which is obviously not optimal and $\check{K} < K^*$ cannot be an optimal solution.

A.6 Proof of Proposition 5

We first prove several supporting lemmas that are useful for the proof of Proposition 5.

Lemma 6.

$$\sum_{t=s}^{s+l-1} \left(\epsilon_t^2 - E\left(\epsilon_t^2\right)\right) = O_P(\sqrt{n\delta_n \ln n}), \tag{A.11}$$

where $1 \leq l \leq n\delta_n$.

Proof. Note $E\left(\epsilon_t^2 - E\left(\epsilon_t^2\right)\right) = 0$ and ϵ_t $1 \leq t \leq n$ are i.i.d. ,

$$\begin{aligned}
&P\left(\left|\sum_{t=s}^{s+l-1}\left(\epsilon_t^2 - E\left(\epsilon_t^2\right)\right)\right| \geq \sqrt{n\delta_n \ln n}\right) \\
&\leq \frac{1}{n\delta_n \ln n} Var\left(\sum_{t=s}^{s+l-1}\epsilon_t^2\right) \\
&= \frac{1}{n\delta_n \ln n}\sum_{t=s}^{s+l-1} 2\sigma^4 \leq \frac{2\sigma^4}{\ln n} \to 0,
\end{aligned}$$

as $n \to \infty$. □

Remark 17. *Since* $\frac{\ln n}{n\delta_n} \to 0$ *as* $n \to \infty$*, so* $\frac{1}{n}\sum_{t=s}^{s+l-1}\epsilon_t^2 = O_P(\delta_n)$ *for* $1 \leq l \leq n\delta_n$*.*

Lemma 7. *If A1-A4 hold, then*

$$\max_{\substack{1\leq r<s\leq n+1\\ s-r\geq n\delta_n}}\left\|\frac{1}{\sqrt{s-r}}\sum_{t=r}^{s-1}\mathbf{x}_t\epsilon_t\right\|_2 = o_P(\sqrt{n\delta_n}J_{min}), \tag{A.12}$$

Proof. For a given constant $c > 0$,

$$\begin{aligned}
&P\left(\max_{\substack{1\leq r<s\leq n+1\\ s-r\geq n\delta_n}}\left\|\frac{1}{\sqrt{s-r}}\sum_{t=r}^{s-1}\mathbf{x}_t\epsilon_t\right\|_2 \geq c\sqrt{n\delta_n}J_{min}\right) \\
&\leq \sum_{\substack{1\leq r<s\leq n+1\\ s-r\geq n\delta_n}} P\left(\left\|\frac{1}{\sqrt{s-r}}\sum_{t=r}^{s-1}\mathbf{x}_t\epsilon_t\right\|_2 \geq c\sqrt{n\delta_n}J_{min}\right) \\
&\leq \sum_{\substack{1\leq r<s\leq n+1\\ s-r\geq n\delta_n}} P\left(\left\|\frac{1}{\sqrt{s-r}}\sum_{t=r}^{s-1}\mathbf{x}_t\epsilon_t\right\|_1 \geq c\sqrt{n\delta_n}J_{min}\right).
\end{aligned}$$

Define $Q_m = \frac{1}{\sqrt{s-r}} \sum_{t=r}^{s-1} x_{t,m}\epsilon_t$ for $1 \le m \le p$, so $Q_m \sim \mathcal{N}\left(0, \frac{\sigma^2}{s-r} \sum_{t=r}^{s-1} x_{t,m}^2\right)$.

$$\begin{aligned}
&\sum_{\substack{1 \le r < s \le n+1 \\ s-r \ge n\delta_n}} P\left(\sum_{m=1}^{p} |Q_m| \ge c\sqrt{n\delta_n} J_{min}\right) \\
&\le \sum_{\substack{1 \le r < s \le n+1 \\ s-r \ge n\delta_n}} P\left(\max_{1 \le m \le p} |Q_m| \ge \frac{c\sqrt{n\delta_n} J_{min}}{p}\right) \\
&\le \sum_{\substack{1 \le r < s \le n+1 \\ s-r \ge n\delta_n}} \sum_{m=1}^{p} P\left(|Q_m| \ge \frac{c\sqrt{n\delta_n} J_{min}}{p}\right) \\
&\le \sum_{\substack{1 \le r < s \le n+1 \\ s-r \ge n\delta_n}} \sum_{m=1}^{p} \exp\left(-\frac{(cJ_{min}\sqrt{n\delta_n}/p)^2}{\frac{2\sigma^2}{s-r}\sum_{t=r}^{s-1} x_{t,m}^2}\right).
\end{aligned}$$

When A1 is satisfied, then

$$\begin{aligned}
\frac{1}{r-s} \sum_{t=s}^{r-1} x_{t,m}^2 &= tr\left(\frac{1}{r-s} \sum_{t=s}^{r-1} \mathbf{x}_t \mathbf{x}_t^T\right) \\
&\le p\mu_{max}\left(\frac{1}{r-s} \sum_{t=s}^{r-1} \mathbf{x}_t \mathbf{x}_t^T\right) \\
&\le pL,
\end{aligned}$$

for $1 \le r < s \le n+1$, $s - r \ge n\delta_n$. From A4 we know J_{min} is lower and upper bounded by some constants. So we have

$$\begin{aligned}
&\sum_{\substack{1 \le r < s \le n+1 \\ s-r \ge n\delta_n}} \sum_{m=1}^{p} \exp\left(-\frac{(cJ_{min}\sqrt{n\delta_n}/p)^2}{\frac{2\sigma^2}{s-r}\sum_{t=r}^{s-1} x_{t,m}^2}\right) \\
&\le \sum_{\substack{1 \le r < s \le n+1 \\ s-r \ge n\delta_n}} \sum_{m=1}^{p} \exp\left(-c^2 c_0 n\delta_n\right) \\
&= p\exp\left(2\ln n - c^2 c_0 n\delta_n\right) \to 0,
\end{aligned}$$

where c_0 is some constant, and the last step is due to $\frac{\ln n}{n\delta_n} \to 0$ as $n \to 0$. □

Lemma 8. *Under A1-A4, If $\hat{K} < K^*$, then for any $\hat{T}_{\hat{K}} = \{\hat{t}_1, \hat{t}_2, \cdots, \hat{t}_{\hat{K}}\}$,*

$$\frac{1}{n}\sum_{k=1}^{\hat{K}+1}\sum_{t=\hat{t}_{k-1}}^{\hat{t}_k-1}(y_t - \hat{\boldsymbol{\alpha}}_k^T\mathbf{x}_t)^2 - \frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2 > c, \tag{A.13}$$

as $n \to \infty$, where c is some positive constant.

Proof. We have $B(\hat{T}_{\hat{K}}) - B(T^*_{K^*}) = \frac{1}{n}\sum_{k=1}^{\hat{K}+1}\sum_{t=\hat{t}_{k-1}}^{\hat{t}_k-1}(y_t - \hat{\boldsymbol{\alpha}}_k^T\mathbf{x}_t)^2 - \frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2$.

Notice that $\hat{\boldsymbol{\alpha}}_k$'s that minimize $B(\hat{T}_{\hat{K}}) - B(T^*_{K^*})$ are the ordinary least squares (OLS) estimators

$$\left(\sum_{t=\hat{t}_{k-1}}^{\hat{t}_k-1}\mathbf{x}_t\mathbf{x}_t^T\right)\hat{\boldsymbol{\alpha}}_k = \sum_{t=\hat{t}_{k-1}}^{\hat{t}_k-1}\mathbf{x}_t y_t.$$

$$\begin{aligned} & B(\hat{T}_{\hat{K}}) - B(T^*_{K^*}) \\ = \ & \frac{1}{n}\sum_{k=1}^{\hat{K}+1}\sum_{t=\hat{t}_{k-1}}^{\hat{t}_k-1}\big[(\boldsymbol{\beta}_t - \hat{\boldsymbol{\alpha}}_k)^T\mathbf{x}_t\mathbf{x}_t^T(\boldsymbol{\beta}_t - \hat{\boldsymbol{\alpha}}_k) \\ & +2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\alpha}}_k)^T\mathbf{x}_t\epsilon_t\big] \equiv \frac{1}{n}\sum_{k=1}^{\hat{K}+1}F_k \end{aligned}$$

Then we show the growth order of F_k in different cases. For simplicity, we show the whole proof for the case $\hat{t}_k - \hat{t}_{k-1} < n\delta_n$ and $[\hat{t}_{k-1}, \hat{t}_k - 1]$ does not contain any true change-point, as other cases are similar.

A.6.1 If $\hat{t}_k - \hat{t}_{k-1} < n\delta_n$

$[\hat{t}_{k-1}, \hat{t}_k - 1]$ **does not contain any true change-point, say it is in some true interval** $[t^*_{l-1}, t^*_l - 1]$

We have

$$\begin{aligned}
\frac{1}{n}F_k &= \frac{1}{n}\sum_{t=\hat{t}_{k-1}}^{\hat{t}_k-1}(y_t - \hat{\boldsymbol{\alpha}}_k^T\mathbf{x}_t)^2 - \frac{1}{n}\sum_{t=\hat{t}_{k-1}}^{\hat{t}_k-1}\epsilon_t^2 \\
&\leq \frac{1}{n}\sum_{t=\hat{t}_{k-1}}^{\hat{t}_k-1}(y_t - \boldsymbol{\alpha}_l^{*T}\mathbf{x}_t)^2 - \frac{1}{n}\sum_{t=\hat{t}_{k-1}}^{\hat{t}_k-1}\epsilon_t^2 \\
&= 0,
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{n}F_k &= \frac{1}{n}\sum_{t=\hat{t}_{k-1}}^{\hat{t}_k-1}(y_t - \hat{\boldsymbol{\alpha}}_k^T\mathbf{x}_t)^2 - \frac{1}{n}\sum_{t=\hat{t}_{k-1}}^{\hat{t}_k-1}\epsilon_t^2 \\
&\geq -\frac{1}{n}\sum_{t=\hat{t}_{k-1}}^{\hat{t}_k-1}\epsilon_t^2 = O_p(\delta_n).
\end{aligned}$$

So $\frac{1}{n}F_k = O_p(\delta_n)$.

If $[\hat{t}_{k-1}, \hat{t}_k - 1]$ **contains some true change-points**

$\frac{1}{n}F_k = O_P(\delta_n)$.

A.6.2 If $\hat{t}_k - \hat{t}_{k-1} \geq n\delta_n$

If $[\hat{t}_{k-1}, \hat{t}_k - 1]$ **does not contain true change-points, say it is in some true interval** $[t^*_{l-1}, t^*_l - 1]$

$\frac{1}{n}F_k = O_P(\delta_n)$.

If $[\hat{t}_{k-1}, \hat{t}_k - 1]$ contains one true change-point, say t_l^*

$t_l^* - \hat{t}_{k-1} < n\delta_n$ **and** $\hat{t}_k - t_l^* < n\delta_n$ $\quad \frac{1}{n}F_k = O_P(\delta_n)$.

$t_l^* - \hat{t}_{k-1} \geq n\delta_n$ **and** $\hat{t}_k - t_l^* < n\delta_n$ $\quad \frac{1}{n}F_k = O_P(\delta_n)$.

$t_l^* - \hat{t}_{k-1} < n\delta_n$ **and** $\hat{t}_k - t_l^* \geq n\delta_n$ $\quad \frac{1}{n}F_k = O_P(\delta_n)$.

$t_l^* - \hat{t}_{k-1} \geq n\delta_n$ **and** $\hat{t}_k - t_l^* \geq n\delta_n$ If $I_l = \Theta(n)$ and $I_{l+1} = \Theta(n)$, then $\frac{1}{n}F_k > c$, where c is some positive constant, otherwise, $\frac{1}{n}F_k = o_P(1)$, where $I_l = t_l^* - \hat{t}_{k-1}$, $I_{l+1} = \hat{t}_k - t_l^*$.

If $[\hat{t}_{k-1}, \hat{t}_k - 1]$ contains two true change-points, say it contains true change-points, say t_l^* and t_{l+1}^*

$t_l^* - \hat{t}_{k-1} < n\delta_n$ **and** $\hat{t}_k - t_{l+1}^* < n\delta_n$ $\quad \frac{1}{n}F_k = O_P(\delta_n)$.

$t_l^* - \hat{t}_{k-1} \geq n\delta_n$ **and** $\hat{t}_k - t_{l+1}^* \geq n\delta_n$ Define $I_{l+2} = \hat{t}_k - t_{l+1}^*$, $I_{l+1} = t_{l+1}^* - t_l^*$ and $I_l = t_l^* - \hat{t}_{k-1}$.

If $I_l = o(n)$ and $I_{l+2} = o(n)$, then $F_k = o_P(1)$.

If $I_l = \Theta(n)$ and $I_{l+2} = \Theta(n)$, then $F_k \geq c + o_P(1)$, where c is some positive constant.

If $I_l = \Theta(n)$ and $I_{l+2} = o(n)$, $F_k \geq c + o_P(1)$, where c is some positive constant.

If $I_l = o(n)$ and $I_{l+2} = \Theta(n)$, then $F_k \geq c + o_P(1)$, where c is some positive constant.

$t_l^* - \hat{t}_{k-1} \geq n\delta_n$ **and** $\hat{t}_k - t_{l+1}^* < n\delta_n$ If $I_l = o(n)$, $\frac{1}{n}F_k = o_P(1)$.

If $I_l = \Theta(n)$, then $\frac{1}{n}F_k > c$.

$t_l^* - \hat{t}_{k-1} < n\delta_n$ **and** $\hat{t}_k - t_{l+1}^* \geq n\delta_n$ If $I_l = o(n)$, then $\frac{1}{n}F_k = o_P(1)$.

If $I_l = \Theta(n)$, then $\frac{1}{n}F_k \geq c + o_P(1)$, where c is some positive constant.

If $[\hat{t}_{k-1}, \hat{t}_k - 1]$ **contains more than two true change-points, say it contains true change- points, say** t_l^***,** $\cdots$**,** t_{l+m}^***,** $m \in \{2, 3, \cdots, K^* - l\}$

$\frac{1}{n} F_k \geq c J_{max} + o_P(1)$, where c is some positive constant.

Then we can discuss for the case $\hat{K} < K^*$, since there is no estimate change-points in the range $\frac{I_{min}}{2}$ of some true change-point, then from the combination of all the cases above, we conclude that $\frac{1}{n} \sum_{k=1}^{\hat{K}+1} \sum_{t=\hat{t}_{k-1}}^{\hat{t}_k - 1} (y_t - \hat{\boldsymbol{\alpha}}_k^T \mathbf{x}_t)^2 - \frac{1}{n} \sum_{t=1}^{n} \epsilon_t^2 > c$, where c is some positive constant. □

Lemma 9. *Under A1-A4, for* $\hat{K} > K^*$*,*

$$B(\hat{T}_{\hat{K}}) - B(\hat{T}_{K^*}) = O_P(\delta_n).$$

Proof. Notice $B(T_{K^*}^*) = \frac{1}{n} \sum_{t=1}^{n} \epsilon_t^2$. We have $|B(\hat{T}_{\hat{K}}) - B(\hat{T}_{K^*})| \leq |B(\hat{T}_{\hat{K}}) - B(T_{K^*}^*)| + |B(\hat{T}_{K^*}) - B(T_{K^*}^*)|$. Since $\hat{K} > K^*$, from Proposition 3, we have $P(\varepsilon(\hat{T}_{\hat{K}} || T^*) \leq n\delta_n) \to 1$, as $n \to \infty$, from the proof of Lemma 8, we have $|B(\hat{T}_{\hat{K}}) - B(T_{K^*}^*)| = O_P(\delta_n)$. From Proposition 1 and the proof of Lemma 8 , we have $|B(\hat{T}_{K^*}) - B(T_{K^*}^*)| = O_P(\delta_n)$. So $B(\hat{T}_{\hat{K}}) - B(\hat{T}_{K^*}) = O_P(\delta_n)$. □

Using the above lemmas, the proof for Proposition 5 follows the same approach of Theorem 3.4 in [40]. We also notice that $B(T_{K^*}^*) = \frac{1}{n} \sum_{t=1}^{n} \epsilon_t^2 \xrightarrow{P} \sigma^2$.

1) $\lambda_n \in \Omega_-$

For $\hat{K} < K^*$, $B(\hat{T}_{K^*}) = B(T_{K^*}^*) + O_P(\delta_n) \xrightarrow{P} \sigma^2 + O_P(\delta_n)$, and $B(\hat{T}_{\hat{K}}) - B(T_{K^*}^*) > c$, where c is some constant,

$$\begin{aligned} & P\left(C(\lambda) > C(\lambda_{K^*})\right) \\ = \; & P\left(\ln \frac{B(\hat{T}_{\hat{K}})}{B(\hat{T}_{K^*})} + \rho_n p(\hat{K} - K^*) > 0\right) \\ = \; & P\left(\ln \frac{B(\hat{T}_{\hat{K}})}{B(\hat{T}_{K^*})} + o_P(1) > 0\right) \to 1, \end{aligned}$$

as $n \to \infty$.

2) $\lambda_n \in \Omega_+$

For $\hat{K} > K^*$, since $\ln(1+x) = x + O(x^2)$, we have

$$\begin{aligned} & \delta_n^{-1}\left(C(\lambda) - C(\lambda_{K^*})\right) \\ = \; & B(\hat{T}_{K^*})^{-1}\delta_n^{-1}\left(B(\hat{T}_{\hat{K}}) - B(\hat{T}_{K^*})\right) \\ & + \delta_n^{-1}\rho_n p(\hat{K} - K^*) + o_P(1). \end{aligned}$$

Note that $\delta_n^{-1}\left(B(\hat{T}_{\hat{K}}) - B(\hat{T}_{K^*})\right) = O_P(1)$ by Lemma 9, $B(\hat{T}_{K^*}) = \sigma^2 + o_P(1)$, and $\frac{\rho_n}{\delta_n} \to \infty$ by Assumption 5, so we have, as $n \to \infty$,

$$\begin{aligned} & P\left(C(\lambda) > C(\lambda_{K^*})\right) \\ = \; & P\left(\delta_n^{-1}\left(\ln \frac{B(\hat{T}_{\hat{K}})}{B(\hat{T}_{K^*})} + \rho_n p(\hat{K} - K^*)\right) > 0\right) \to 1. \end{aligned}$$

A.7 Proof of Proposition 6

A.7.1 Inner loop

It can be seen from Algorithm 2 and Procedure 3 in [31] that the complexity of inner loop is $O(n)$.

A.7.2 Outer loop

Since the solution of the inner loop is exact and the outer loop is randomized, SSLS is a randomized block-coordinate descent method for minimizing a composite function, whose complexity has been recently analyzed in Theorem 5 of [74]. For any given initial value $\boldsymbol{\theta}_0$, the complexity is related to $\phi(\boldsymbol{\theta}_0)$ and $R_2(\boldsymbol{\theta}_0) = \max_{\boldsymbol{\theta}}\{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2 : \phi(\boldsymbol{\theta}) \leq$

$\phi(\boldsymbol{\theta}_0)\}$,which is a measure of the size of the level set of ϕ given by $\boldsymbol{\theta}_0$. In SSLS, we have $\boldsymbol{\theta}_0 = \mathbf{0}$. In the following, we provides some upper bounds on these two quantities, and then apply the results in [74] to provide a bound on the complexity of SSLS.

Firstly, we bound $\phi(\boldsymbol{\theta}_0)$. We have

$$\begin{aligned}
&\phi(\boldsymbol{\theta}_0) = \sum_{k=1}^{K^*+1} \frac{t_k^* - t_{k-1}^*}{n} \boldsymbol{\alpha}_k^{*T} \left(\frac{1}{t_k^* - t_{k-1}^*} \sum_{t=t_{k-1}^*}^{t_k^*-1} \mathbf{x}_t \mathbf{x}_t^T \right) \boldsymbol{\alpha}_k^* \\
&+ \sum_{k=1}^{K^*+1} \frac{t_i^* - t_{k-1}^*}{n} \boldsymbol{\alpha}_i^{*T} \left(\frac{1}{t_k^* - t_{k-1}^*} \sum_{t=t_{k-1}^*}^{t_k^*-1} \mathbf{x}_t \epsilon_t \right) + \frac{1}{n} \sum_{t=1}^{n} \epsilon_t^2 \\
&\leq \sum_{k=1}^{K^*+1} \frac{t_k^* - t_{k-1}^*}{n} L \|\boldsymbol{\alpha}_k^*\|_2^2 + \sum_{k=1}^{K^*+1} \|\boldsymbol{\alpha}_k^*\|_2 o_P(J_{min}) \\
&+ O_P(1) + \sigma^2 \leq M,
\end{aligned}$$

with probability goes to 1 when $n \to \infty$, where M is a constant independent of n.

Secondly, we bound $R_2(\boldsymbol{\theta}_0)$. For those $\boldsymbol{\theta}$s with $\phi(\boldsymbol{\theta}) \leq \phi(\boldsymbol{\theta}_0)$, we can show

$$\lambda_n \sum_{t=1}^{n} \|\boldsymbol{\theta}_t\|_2 \leq \frac{1}{n} \|\mathbf{Y}\|_2^2, \|\boldsymbol{\theta}\|_2 \leq \frac{M}{\lambda_n},$$

with probability goes to 1 when $n \to \infty$. Hence, $\left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right\|_2 \leq 2\frac{M}{\lambda_n}$, which implies $R_2(\boldsymbol{\theta}_0) \leq 2\frac{M}{\lambda_n}$. Let ϕ^* denote the optimal value of function ϕ, and $\phi^* < \phi(\boldsymbol{\theta}^*) \xrightarrow{P} \sigma^2$, as $n \to \infty$.

For any given confidence level $0 < \rho < 1$ and error tolerance $\epsilon > 0$, we then plug bounds obtained above into Theorem 5 in [74] to complete the proof.

A.8 Proof of Supporting Lemmas in Section 3.2.2

A.8.1 Proof of Lemma 1

Since $\hat{\boldsymbol{\theta}}$ is the optimal solution to the optimization problem (3.1), we have

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

which implies

$$\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^*) + \lambda_n \mathcal{R}(\hat{\boldsymbol{\theta}}) \leq \lambda_n \mathcal{R}(\boldsymbol{\theta}^*). \tag{A.14}$$

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

$$\begin{aligned} & \mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^*) \\ = \; & \frac{1}{n}\|\mathbf{y} - \tilde{\mathbf{X}}(\boldsymbol{\theta}^* + \boldsymbol{\Delta}_{\boldsymbol{\theta}})\|_2^2 - \frac{1}{n}\|\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\theta}^*\|_2^2 \\ = \; & \frac{1}{n}\boldsymbol{\Delta}_{\boldsymbol{\theta}}^T \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \boldsymbol{\Delta}_{\boldsymbol{\theta}} - \frac{2}{n}\mathbf{e}^T \tilde{\mathbf{X}} \boldsymbol{\Delta}_{\boldsymbol{\theta}} \\ = \; & \frac{1}{n}\boldsymbol{\Delta}_{\boldsymbol{\theta}}^T \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \boldsymbol{\Delta}_{\boldsymbol{\theta}} + \langle \nabla \mathcal{L}(\boldsymbol{\theta}^*), \boldsymbol{\Delta}_{\boldsymbol{\theta}} \rangle. \end{aligned} \tag{A.15}$$

In (A.15), the first term is related to the data and estimation error, the second term is also related to the noise vector $\mathbf{e}$. For the second term, if we choose $\lambda_n \geq 2\mathcal{R}^* \left(\nabla \mathcal{L}(\boldsymbol{\theta}^*)\right)$, we have

$$\begin{aligned} |\langle \nabla \mathcal{L}(\boldsymbol{\theta}^*), \boldsymbol{\Delta}_{\boldsymbol{\theta}} \rangle| & \overset{(a)}{\leq} \mathcal{R}^*(\nabla \mathcal{L}(\boldsymbol{\theta}^*)) \mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) \\ & \overset{(b)}{\leq} \frac{\lambda_n}{2} \mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}), \end{aligned} \tag{A.16}$$

in which (a) holds due to the definition of the dual norm, and (b) is true due to the fact that we choose $\lambda_n \geq 2\mathcal{R}^* \left(\nabla \mathcal{L}(\boldsymbol{\theta}^*)\right)$.

Plugging (A.16) in (A.15), we have

$$\frac{1}{n}\boldsymbol{\Delta}_{\boldsymbol{\theta}}^T\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\boldsymbol{\Delta}_{\boldsymbol{\theta}} - \frac{\lambda_n}{2}\mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) \leq \mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^*) \leq \frac{1}{n}\boldsymbol{\Delta}_{\boldsymbol{\theta}}^T\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\boldsymbol{\Delta}_{\boldsymbol{\theta}} + \frac{\lambda_n}{2}\mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}). \quad \text{(A.17)}$$

Using (A.17), we have

$$\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^*) \geq \frac{1}{n}\boldsymbol{\Delta}_{\boldsymbol{\theta}}^T\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\boldsymbol{\Delta}_{\boldsymbol{\theta}} - \frac{\lambda_n}{2}\mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) \geq -\frac{\lambda_n}{2}\mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}). \quad \text{(A.18)}$$

Plugging (A.18) into (A.14), we have

$$-\frac{\lambda_n}{2}\mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) + \lambda_n\mathcal{R}(\boldsymbol{\theta}^* + \boldsymbol{\Delta}_{\boldsymbol{\theta}}) \;\; \leq \;\; \lambda_n\mathcal{R}(\boldsymbol{\theta}^*),$$

which coupled with the triangle inequality leads to

$$-\frac{1}{2}\mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) + \mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^*) \;\; \leq \;\; \mathcal{R}(\boldsymbol{\theta}^*).$$

This is the same as

$$\mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) \;\; \leq \;\; 4\mathcal{R}(\boldsymbol{\theta}^*).$$

A.8.2 Proof of Lemma2

If $0 < l_{\mathbf{U}} \leq \mu_{\min}(\boldsymbol{\Sigma}_{\mathbf{U}}) \leq \mu_{\max}(\boldsymbol{\Sigma}_{\mathbf{U}}) \leq L_{\mathbf{U}} < \infty$, we have

$$\|\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{v}\|_2^2 = \boldsymbol{v}^T{\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}}^T\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{v} \geq \mu_{\min}(\boldsymbol{\Sigma}_{\mathbf{U}})\|\boldsymbol{v}\|_2^2 \geq l_{\mathbf{U}}\|\boldsymbol{v}\|_2^2,$$

and

$$\|\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{v}\|_2^2 = \boldsymbol{v}^T{\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}}^T\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{v} \leq \mu_{\max}(\boldsymbol{\Sigma}_{\mathbf{U}})\|\boldsymbol{v}\|_2 \leq L_{\mathbf{U}}\|\boldsymbol{v}\|_2^2,$$

which completes the proof for (3.7).

From [47], for all $\boldsymbol{v} \in \mathbb{R}^{m_2}$ we have

$$\frac{1}{\sqrt{m_1}}\|\mathbf{U}\boldsymbol{v}\|_2 \quad \geq \quad \frac{1}{4}\|\boldsymbol{\Sigma}_{\mathbf{U}}^{\frac{1}{2}}\boldsymbol{v}\|_2 - 9\rho(\boldsymbol{\Sigma}_{\mathbf{U}})\sqrt{\frac{\log m_2}{m_1}}\|\boldsymbol{v}\|_1$$

with probability at least $1 - c^{'} \exp\left(-cm_1\right)$.

Then using the inequality $(a-b)^2 \geq \frac{1}{2}a^2 - ab$ along with (3.7), we obtain (3.8).

A.9 Proof for Consistency Results in Section 3.2.2

Define

$$\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) := \varphi(\boldsymbol{\theta}^* + \boldsymbol{\Delta}_{\boldsymbol{\theta}}) - \varphi(\boldsymbol{\theta}^*).$$

For the optimal solution $\hat{\boldsymbol{\theta}}$ to (3.1), we must have $\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) \leq 0$. The main idea of our proof is to analyze $\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}})$ to identify those $\boldsymbol{\Delta}_{\boldsymbol{\theta}}$'s such that $\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) > 0$. In particular, we show that if the conclusions in the propositions do not hold, we will have $\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) > 0$. In other words, for $\hat{\boldsymbol{\theta}}$ to be the optimal solution to (3.1), the conclusions in the propositions must hold.

A.9.1 Proof for Proposition 7

We have

$$
\begin{aligned}
\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) &= \mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^*) + \lambda_n \left(\mathcal{R}(\hat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^*)\right) \\
&\overset{(a)}{\geq} \frac{1}{n}\left\|\tilde{\mathbf{X}}\boldsymbol{\Delta}_{\boldsymbol{\theta}}\right\|_2^2 - \frac{1}{2}\lambda_n \mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) + \lambda_n \left(\mathcal{R}(\hat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^*)\right) \\
&\geq \frac{1}{n}\left\|\tilde{\mathbf{X}}\boldsymbol{\Delta}_{\boldsymbol{\theta}}\right\|_2^2 - \lambda_n \mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) + \lambda_n \left(\mathcal{R}(\hat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^*)\right) \\
&\overset{(b)}{\geq} \frac{1}{n}\left\|\mathbf{X}\boldsymbol{\Delta}_{\boldsymbol{\beta}}\right\|_2^2 - 2\lambda_n \mathcal{R}(\boldsymbol{\theta}^*), \qquad \text{(A.19)}
\end{aligned}
$$

in which (a) comes from (A.17), and (b) is due to (3.5) and the triangle inequality.

Define events

$$
\begin{aligned}
A_{n,j} &:= \left\{|\hat{t}_j - t_j^*| > n\delta_n\right\}, \\
C_n &:= \left\{\max_{1 \leq k \leq K^*} |\hat{t}_k - t_k^*| < I_{\min}/2\right\}.
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&\mathbb{P}\left\{\max_{1 \leq k \leq K^*} |\hat{t}_k - t_k^*| \leq n\delta_n\right\} \\
\geq\ & 1 - \sum_{j=1}^{K^*} \mathbb{P}\left\{A_{n,j}\right\} \\
=\ & 1 - \sum_{j=1}^{K^*} \mathbb{P}\left\{A_{n,j} \cap C_n\right\} - \sum_{j=1}^{K^*} \mathbb{P}\left\{A_{n,j} \cap C_n^c\right\}.
\end{aligned}
$$

To prove Proposition 7, 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.

Bounding $\mathbb{P}\left\{A_{n,j} \cap C_n\right\}$

Define

$$\begin{aligned} A_{n,j}^{+} &:= \{t_j^* - \hat{t}_j > n\delta_n\}, \\ A_{n,j}^{-} &:= \{\hat{t}_j - t_j^* > n\delta_n\}. \end{aligned}$$

It suffices to prove our results under $A_{n,j} \cap C_n$ by showing under the cases $A_{n,j}^{+} \cap C_n$ and $A_{n,j}^{-} \cap C_n$.

We first prove under the case $A_{n,j}^{+} \cap C_n$ and the other case follows similarly. On $A_{n,j}^{+} \cap C_n$, we have $t_{k-1}^* < \hat{t}_k < t_{k+1}^*$ for all $k = 1, \cdots, K^*$. For j we have, $t_j^* - \hat{t}_j > n\delta_n$ and $n > \hat{t}_{j+1} - t_j^* > \frac{I_{\min}}{2}$. Here we prove the case $\hat{t}_{j+1} \leq t_{j+1}^*$ and the other case follows similarly. Now we turn to bound $\mathbb{P}\left\{A_{n,j}^{+} \cap C_n \cap \{\hat{t}_{j+1} \leq t_{j+1}^*\}\right\}$. The case under study is illustrated in Figure A.3.

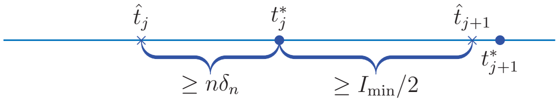

Figure A.3: Illustration of the case $t_j^* - \hat{t}_j > n\delta_n$, $n > \hat{t}_{j+1} - t_j^* > \frac{I_{\min}}{2}$ and $\hat{t}_{j+1} \leq t_{j+1}^*$.

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 (A.19), we have

$$\begin{aligned}
\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) &\geq \frac{1}{n}\|\mathbf{X}\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2^2 - 2\lambda_n \mathcal{R}(\boldsymbol{\theta}^*) \\
&\geq \frac{1}{n}\sum_{t=\hat{t}_j}^{\hat{t}_{j+1}-1}(\mathbf{x}_t^T(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^*))^2 - 2\lambda_n \mathcal{R}(\boldsymbol{\theta}^*) \\
&= \frac{1}{n}\left\|\begin{pmatrix} \mathbf{x}_{\hat{t}_j}^T \\ \cdots \\ \mathbf{x}_{t_j^*-1}^T \end{pmatrix}(\boldsymbol{\alpha}_j^* - \hat{\boldsymbol{\alpha}}_{j+1})\right\|_2^2 \\
&\quad + \frac{1}{n}\left\|\begin{pmatrix} \mathbf{x}_{t_j^*}^T \\ \cdots \\ \mathbf{x}_{\hat{t}_{j+1}-1}^T \end{pmatrix}(\boldsymbol{\alpha}_{j+1}^* - \hat{\boldsymbol{\alpha}}_{j+1})\right\|_2^2 - 2\lambda_n \mathcal{R}(\boldsymbol{\theta}^*). \qquad \text{(A.20)}
\end{aligned}$$

Using Lemma 2, we have

$$\begin{aligned}
&\frac{1}{n}\left\|\begin{pmatrix} \mathbf{x}_{\hat{t}_j}^T \\ \cdots \\ \mathbf{x}_{t_j^*-1}^T \end{pmatrix}(\boldsymbol{\alpha}_j^* - \hat{\boldsymbol{\alpha}}_{j+1})\right\|_2^2 \\
\overset{(a)}{\geq}\ & \frac{t_j^* - \hat{t}_j}{n}\left(\frac{l}{32}\|\boldsymbol{\alpha}_j^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_2^2 - \frac{9}{4}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{t_j^* - \hat{t}_j}}\|\boldsymbol{\alpha}_j^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_1\|\boldsymbol{\alpha}_j^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_2\right) \\
\overset{(b)}{>}\ & \frac{l}{32}\delta_n\|\boldsymbol{\alpha}_j^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_2^2 - \frac{9}{4}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{t_j^* - \hat{t}_j}{n}}\sqrt{\frac{\log p}{n}}\|\boldsymbol{\alpha}_j^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_1\|\boldsymbol{\alpha}_j^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_2 \\
\geq\ & \frac{l}{32}\delta_n\|\boldsymbol{\alpha}_j^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_2^2 - \frac{9}{4}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\|\boldsymbol{\alpha}_j^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_1\|\boldsymbol{\alpha}_j^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_2, \qquad \text{(A.21)}
\end{aligned}$$

with probability at least $1 - c^{'}\exp\left(-cn\delta_n\right)$ and this probability is from (a), in which we uses Lemma 2. Step (b) uses the fact that $t_j^* - \hat{t}_j \geq n\delta_n$.

Similar results also hold for the second quadratic term in (A.20). In particular, as

$n > \hat{t}_{j+1} - t_j^* > \frac{I_{\min}}{2}$, following similar steps as in (A.21), we have

$$\begin{aligned} & \frac{1}{n}\left\| \begin{pmatrix} \mathbf{x}_{t_j^*}^T \\ \cdots \\ \mathbf{x}_{\hat{t}_{j+1}-1}^T \end{pmatrix} (\boldsymbol{\alpha}_{j+1}^* - \hat{\boldsymbol{\alpha}}_{j+1}) \right\|_2^2 \\ > \; & \frac{l}{32}\frac{I_{\min}}{2n}\|\boldsymbol{\alpha}_{j+1}^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_2^2 \\ & -\frac{9}{4}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\|\boldsymbol{\alpha}_{j+1}^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_1\|\boldsymbol{\alpha}_{j+1}^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_2. \qquad \text{(A.22)} \end{aligned}$$

with probability at least $1 - c^{'} \exp(-cI_{\min}/2)$.

From the definition of $\mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}})$, we have

$$\begin{aligned} \frac{1}{1-\gamma}\mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) & \geq \|\boldsymbol{\Delta}_{\boldsymbol{\theta}}\|_1, \qquad \text{(A.23)} \\ & \geq \sum_{t=1}^{\hat{t}_j}\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t^*\|_1 \\ & = \sum_{t=1}^{\hat{t}_j}\|\hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}_t^* + \boldsymbol{\beta}_{t-1}^*\|_1 \\ & \geq \sum_{t=1}^{\hat{t}_j}\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^*\|_1 - \sum_{t=1}^{\hat{t}_j}\|\hat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}_{t-1}^*\|_1 \\ & = \|\hat{\boldsymbol{\alpha}}_{j+1} - \boldsymbol{\alpha}_j^*\|_1. \qquad \text{(A.24)} \end{aligned}$$

Combining this with (3.6), we have

$$\|\hat{\boldsymbol{\alpha}}_{j+1} - \boldsymbol{\alpha}_j^*\|_1 \leq \frac{4}{1-\gamma}\mathcal{R}(\boldsymbol{\theta}^*). \qquad \text{(A.25)}$$

Using the same technique, we also have

$$\|\hat{\boldsymbol{\alpha}}_{j+1} - \boldsymbol{\alpha}_{j+1}^*\|_1 \leq \frac{4}{1-\gamma}\mathcal{R}(\boldsymbol{\theta}^*). \qquad \text{(A.26)}$$

We can obtain similar bounds for the ℓ_2-norm terms.

Plugging (A.21), (A.22), (A.25), (A.26) and corresponding equations for ℓ_2-norm into (A.20), we obtain

$$\begin{aligned}\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) &> \frac{l}{32}\left(\delta_n\|\boldsymbol{\alpha}_j^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_2^2 + \frac{I_{\min}}{2n}\|\boldsymbol{\alpha}_{j+1}^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_2^2\right)\\ &\quad -\frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) - 2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*).\end{aligned}$$

Since $\delta_n \leq \frac{I_{\min}}{2n}$, we can view $\delta_n\|\boldsymbol{\alpha}_j^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_2^2 + \frac{I_{\min}}{2n}\|\boldsymbol{\alpha}_{j+1}^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_2^2$ as a function of $\hat{\boldsymbol{\alpha}}_{j+1}$. Furthermore, this function is convex and we can easily find its minimum to be $\frac{\delta_n}{2}\|\boldsymbol{\alpha}_{j+1}^* - \boldsymbol{\alpha}_j^*\|_2^2$. As the result, we have

$$\begin{aligned}\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) &> \frac{l}{64}\delta_n\|\boldsymbol{\alpha}_{j+1}^* - \boldsymbol{\alpha}_j^*\|_2^2 - \frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) - 2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*)\\ &= 0, \qquad\qquad (A.27)\end{aligned}$$

with probability at least $1 - c^{'}\exp(-cn\delta_n) - c^{'}\exp(-cI_{\min}/2)$, where the last step is due to the definition of δ_n. Hence,

$$\mathbb{P}\left\{\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) > 0 \;\middle|\; A_{n,j}^+ \cap C_n \cap \{\hat{t}_{j+1} \leq t_{j+1}^*\}\right\} \geq 1 - c^{'}\exp(-cn\delta_n) - c^{'}\exp(-cI_{\min}/2).$$

Since $\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) > 0$ cannot happen, hence if $1 - c^{'}\exp(-cn\delta_n) - c^{'}\exp(-cI_{\min}/2) > 0$, then

$$\mathbb{P}\left\{A_{n,j}^+ \cap C_n \cap \{\hat{t}_{j+1} \leq t_{j+1}^*\}\right\} = 0.$$

For $\mathbb{P}\left\{A_{n,j}^+ \cap C_n \cap \{\hat{t}_{j+1} > t_{j+1}^*\}\right\}$, by focusing on the interval $[\hat{t}_j, t_{j+1}^* - 1]$, we use the same technique as above and show

$$\mathbb{P}\left\{\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) > 0 \;\middle|\; A_{n,j}^+ \cap C_n \cap \{\hat{t}_{j+1} > t_{j+1}^*\}\right\} \geq 1 - c^{'}\exp(-cn\delta_n) - c^{'}\exp(-cI_{\min}).$$

Hence $\mathbb{P}\left\{A_{n,j}^{+} \cap C_n \cap \{\hat{t}_{j+1} > t_{j+1}^{*}\}\right\} = 0$ if $1 - c^{'} \exp\left(-cn\delta_n\right) - c^{'} \exp\left(-cI_{\min}\right) > 0$.

For the case $A_{n,j}^{-}$, we turn to study the position of $\hat{t}_{j-1}$ and split this case into either $\hat{t}_{j-1} \geq t_{j-1}^{*}$ or $\hat{t}_{j-1} < t_{j-1}^{*}$, and we have the same result as $A_{n,j}^{+}$ due to symmetry. Hence if $1 - c^{'} \exp\left(-cn\delta_n\right) - c^{'} \exp\left(-cI_{\min}/2\right) > 0$,

$$\mathbb{P}\left\{A_{n,j} \cap C_n\right\} = 0. \tag{A.28}$$

Bounding $\mathbb{P}\left\{A_{n,j} \cap C_n^c\right\}$

Here we prove under the case $A_{n,j}^{+}$ since the other case follow similarly due to symmetry. Define

$$\begin{aligned}
D_T^{(l)} &:= \left\{\exists j \in \{1, \cdots, K^*\}, \hat{t}_j \leq t_{j-1}^{*}\right\} \cap C_n^c, \\
D_T^{(m)} &:= \left\{\forall j \in \{1, \cdots, K^*\}, t_{j-1}^{*} < \hat{t}_j < t_{j+1}^{*}\right\} \cap C_n^c, \\
D_T^{(r)} &:= \left\{\exists j \in \{1, \cdots, K^*\}, \hat{t}_j \geq t_{j+1}^{*}\right\} \cap C_n^c.
\end{aligned}$$

Then we have $A_{n,j}^{+} \cap C_n^c \subseteq \left(A_{n,j}^{+} \cap D_T^{(l)}\right) \cup \left(A_{n,j}^{+} \cap D_T^{(m)}\right) \cup \left(A_{n,j}^{+} \cap D_T^{(r)}\right)$. First we study the case $A_{n,j}^{+} \cap D_T^{(m)}$.

$$\begin{aligned}
& A_{n,j}^{+} \cap D_T^{(m)} \\
= & \left(A_{n,j}^{+} \cap \left\{\hat{t}_{j+1} - t_j^{*} \geq I_{min}/2\right\} \cap D_T^{(m)}\right) \cup \left(A_{n,j}^{+} \cap \left\{\hat{t}_{j+1} - t_j^{*} < I_{min}/2\right\} \cap D_T^{(m)}\right) \\
\subseteq & \left(A_{n,j}^{+} \cap \left\{\hat{t}_{j+1} - t_j^{*} \geq I_{min}/2\right\} \cap D_T^{(m)}\right) \cup \left(A_{n,j}^{+} \cap \left\{t_{j+1}^{*} - \hat{t}_{j+1} \geq I_{min}/2\right\} \cap D_T^{(m)}\right).
\end{aligned}$$

Since

$$\begin{aligned}
& \left\{A_{n,j}^{+} \cap \left\{t_{j+1}^{*} - \hat{t}_{j+1} \geq I_{min}/2\right\} \cap D_T^{(m)}\right\} \\
& \subseteq \cup_{k=j+1}^{K^*} \left(\{t_k^{*} - \hat{t}_k \geq I_{\min}/2\} \cap \{\hat{t}_{k+1} - t_k^{*} \geq I_{\min}/2\} \cap D_T^{(m)}\right),
\end{aligned}$$

we have

$$\begin{aligned}
& \mathbb{P}\left\{A_{n,j}^{+} \cap D_T^{(m)}\right\} \\
\leq\ & \mathbb{P}\left\{A_{n,j}^{+} \cap \{\hat{t}_{j+1} - t_j^* \geq I_{min}/2\} \cap D_T^{(m)}\right\} \\
& + \sum_{k=j+1}^{K^*} \mathbb{P}\left\{\{t_k^* - \hat{t}_k \geq I_{\min}/2\} \cap \{\hat{t}_{k+1} - t_k^* \geq I_{\min}/2\} \cap D_T^{(m)}\right\}. \quad \text{(A.29)}
\end{aligned}$$

We first study $\mathbb{P}\left\{A_{n,j}^{+} \cap \{\hat{t}_{j+1} - t_j^* \geq I_{min}/2\} \cap D_T^{(m)}\right\}$ since other terms follow similarly.

Noticing that the upper bound of each term in (A.29) can be found using the same technique in Appendix A.9.1, we have that

$$\mathbb{P}\left\{A_{n,j}^{+} \cap \{\hat{t}_{j+1} - t_j^* \geq I_{min}/2\} \cap D_T^{(m)}\right\} = 0,$$

if $1 - \left(c^{'} \exp\left(-cn\delta_n\right) + c^{'} \exp\left(-cI_{\min}/2\right)\right) > 0$, and have

$$\mathbb{P}\left\{\{t_{K^*}^* - \hat{t}_{K^*} \geq I_{\min}/2\} \cap \{\hat{t}_{K^*+1} - t_{K^*}^* \geq I_{\min}/2\} \cap D_T^{(m)}\right\} = 0,$$

if $1 - 2c^{'} \exp\left(-cI_{\min}/2\right) > 0$.

Combing them together we have

$$\mathbb{P}\left\{A_{n,j}^{+} \cap D_T^{(m)}\right\} = 0, \text{ if } 1 - \left(c^{'} \exp\left(-cn\delta_n\right) + c^{'} \exp\left(-cI_{\min}/2\right)\right) > 0. \quad \text{(A.30)}$$

Similarly, for the case $A_{n,j}^-$, due to symmetry, we have

$$\begin{aligned}
&\mathbb{P}\left\{A_{n,j}^- \cap D_T^{(m)}\right\} \\
\leq\ & \mathbb{P}\left\{A_{n,j}^- \cap \{t_j^* - \hat{t}_{j-1} \geq I_{min}/2\} \cap D_T^{(m)}\right\} \\
&+ \sum_{k=1}^{j-1} \mathbb{P}\left\{\{\hat{t}_k - t_k^* \geq I_{\min}/2\} \cap \{t_k^* - \hat{t}_{k-1} \geq I_{\min}/2\} \cap D_T^{(m)}\right\} \\
=\ & 0, \qquad \text{(A.31)}
\end{aligned}$$

if $1 - \left(c^{'} \exp\left(-cn\delta_n\right) + c^{'} \exp\left(-cI_{\min}/2\right)\right) > 0$.

Combining (A.30) and (A.31), we have

$$\mathbb{P}\left\{A_{n,j} \cap D_T^{(m)}\right\} = 0, \text{ if } 1 - \left(c^{'} \exp\left(-cn\delta_n\right) + c^{'} \exp\left(-cI_{\min}/2\right)\right) > 0. \qquad \text{(A.32)}$$

Then we turn to find upper bound of $\mathbb{P}\left\{A_{n,j}^+ \cap D_T^{(l)}\right\}$ and $\mathbb{P}\left\{A_{n,j}^+ \cap D_T^{(r)}\right\}$. The method to find upper bounds of them is the same due to symmetry. Since

$$\mathbb{P}\left\{A_{n,j}^+ \cap D_T^{(l)}\right\} \leq \mathbb{P}\left\{D_T^{(l)}\right\} \leq \sum_{j=1}^{K^*} 2^{j-1}\, \mathbb{P}\left\{\max\{1 \leq l \leq K^* | \hat{t}_l \leq t_{j-1}^*\} = j\}\right\},$$

and

$$\left\{\max\{1 \leq l \leq K^* | \hat{t}_l \leq t_{j-1}^*\} = j\right\} \subseteq \cup_{k=j}^{K^*-1} \left(\{t_k^* - \hat{t}_k \geq I_{\min}/2\} \cap \{\hat{t}_{k+1} - t_k^* \geq I_{\min}/2\}\right),$$

then we have

$$\begin{aligned}
\mathbb{P}\left\{A_{n,j}^+ \cap D_T^{(l)}\right\} \ \leq\ & \sum_{j=1}^{K^*-1} 2^{j-1} \sum_{k=j}^{K^*-1} \mathbb{P}\left\{\{t_k^* - \hat{t}_k \geq I_{\min}/2\} \cap \{\hat{t}_{k+1} - t_k^* \geq I_{\min}/2\}\right\} \\
&+ 2^{K^*-1}\, \mathbb{P}\left\{t_{K^*}^* - \hat{t}_{K^*} \geq I_{\min}/2\right\}.
\end{aligned}$$

Following the same reasoning in Subsection A.9.1, we have

$$\mathbb{P}\left\{A_{n,j}^{+} \cap D_T^{(l)}\right\} = 0,\ \text{if}\ 1 - \left(2c^{'} \exp\left(-cI_{\min}/2\right)\right) > 0.$$

Similarly, we have

$$\mathbb{P}\left\{A_{n,j}^{-} \cap D_T^{(l)}\right\} = 0,\ \text{if}\ 1 - \left(2c^{'} \exp\left(-cI_{\min}/2\right)\right) > 0.$$

Combining them together and since $n\delta_n < I_{\min}/2$, we have

$$\mathbb{P}\left\{A_{n,j} \cap D_T^{(l)}\right\} = 0,\ \text{if}\ 1 - \left(c^{'} \exp\left(-cn\delta_n\right) + c^{'} \exp\left(-cI_{\min}/2\right)\right) > 0$$

and

$$\mathbb{P}\left\{A_{n,j} \cap D_T^{(r)}\right\} = 0,\ \text{if}\ 1 - \left(c^{'} \exp\left(-cn\delta_n\right) + c^{'} \exp\left(-cI_{\min}/2\right)\right) > 0.$$

Thus we find an upper bound for $\mathbb{P}\left\{A_{n,j} \cap C_n^c\right\}$,

$$\mathbb{P}\left\{A_{n,j} \cap C_n^c\right\} = 0,\ \text{if}\ 1 - \left(c^{'} \exp\left(-cn\delta_n\right) + c^{'} \exp\left(-cI_{\min}/2\right)\right) > 0. \tag{A.33}$$

Combing (A.28) and (A.33), we obtained an upper bound of $\mathbb{P}\left\{A_{n,j}\right\}$,

$$\mathbb{P}\left\{A_{n,j}\right\} = 0,\ \text{if}\ 1 - \left(c^{'} \exp\left(-cn\delta_n\right) + c^{'} \exp\left(-cI_{\min}/2\right)\right) > 0.$$

Due to the definition of $A_{n,k}$, we have

$$\begin{aligned} & \mathbb{P}\left\{\exists k \in [K^*] \text{ such that } |\hat{t}_k - t_k^*| > n\delta_n\right\} \\ = & \mathbb{P}\left\{\cup_{k=1}^{K^*} A_{n,k}\right\} \\ \leq & \sum_{k=1}^{K^*} \mathbb{P}\left\{A_{n,k}\right\} \\ = & 0, \end{aligned}$$

if $1 - \left(c^{'} \exp\left(-cn\delta_n\right) + c^{'} \exp\left(-cI_{\min}/2\right)\right) > 0$.

And from Assumption B1 we know that

$$1 - \left(c^{'} \exp\left(-cn\delta_n\right) + c^{'} \exp\left(-cI_{\min}/2\right)\right) > 1 - 2c^{'} \exp\left(-cn\delta_n\right) > 0.$$

So we have

$$\max_{1 \leq k \leq K^*} |\hat{t}_k - t_k^*| \leq n\delta_n.$$

A.9.2 Proof for Proposition 8

Since Proposition 7 holds, we have $|\hat{t}_k - t_k^*| \leq n\delta_n$ for each $k = 1, \cdots, K^*$. Then if $\|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\|_2^2 > \frac{32n}{(I_{\min} - 2n\delta_n)l}\left(\frac{36}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) + 2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*)\right)$, we have

$$
\begin{aligned}
&\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) \\
&\stackrel{(a)}{\geq} \frac{1}{n}\|\mathbf{X}\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2^2 - 2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*) \\
&\geq \frac{1}{n}\sum_{t=t_{k-1}^*-1}^{t_k^*}\left(\mathbf{x}_t^T\left(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\alpha}_k^*\right)\right)^2 - 2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*) \\
&\geq \frac{1}{n}\sum_{t=t_{k-1}^*-1+n\delta_n}^{t_k^*-n\delta_n}\left(\mathbf{x}_t^T\left(\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\right)\right)^2 - 2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*) \\
&\stackrel{(b)}{\geq} \frac{m_k}{n}\left(\frac{l}{32}\|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\|_2^2 - \frac{9}{4}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{m_k}}\|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\|_1\|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\|_2\right) \\
&\quad -2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*) &&\text{(A.34)} \\
&\geq \frac{m_k}{n}\frac{l}{32}\|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\|_2^2 - \frac{9}{4}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\|_1\|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\|_2 \\
&\quad -2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*) &&\text{(A.35)} \\
&\stackrel{(c)}{\geq} \frac{I_{\min} - 2n\delta_n}{n}\frac{l}{32}\|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\|_2^2 - \frac{36}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) \\
&\quad -2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*), &&\text{(A.36)} \\
&> 0.
\end{aligned}
$$

in which (a) is due to (A.19), (b) is due to Lemma 2 with $m_k = t_k^* - t_{k-1}^* - 2n\delta_n$ and holds with probability at least $1 - c^{'}\exp\left(-c\left(I_{\min} - 2n\delta_n\right)\right)$, and (c) is due to the definition of I_{min}, (A.23) combined with Lemma 1. Hence we have

$$
\begin{aligned}
&\mathbb{P}\left\{\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) > 0 \;\middle|\; \|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\|_2^2 > \frac{32n}{(I_{\min} - 2n\delta_n)l}\left(\frac{36}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) + 2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*)\right)\right\} \\
&\geq \;\; 1 - c^{'}\exp\left(-c\left(I_{\min} - 2n\delta_n\right)\right) > 0,
\end{aligned}
$$

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

$$\|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\|_2^2 \leq \frac{32n}{(I_{\min} - 2n\delta_n)l} \left(\frac{36}{(1-\gamma)^2} \sqrt{L}\rho(\boldsymbol{\Sigma}) \sqrt{\frac{\log p}{n}} \mathcal{R}^2(\boldsymbol{\theta}^*) + 2\lambda_n \mathcal{R}(\boldsymbol{\theta}^*) \right).$$

A.9.3 Proof for Proposition 9

Suppose $\hat{K} < K^*$, then there exists at least one true change-point t_k^* such that there is no estimated change-points in the range of $I_{\min}/2$ around it. This scenario is illustrated in Figure A.4. Following the same proof as Appendix A.9.1, we have

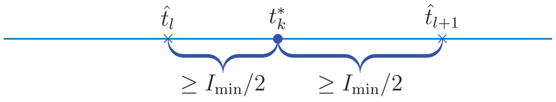

Figure A.4: Illustration of the case with $\hat{t}_{l+1} - t_k^* \geq I_{\min}/2$ and $t_k^* - \hat{t}_l \geq I_{\min}/2$.

$$\begin{aligned}
\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) &\geq \frac{1}{n}\|\mathbf{X}\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2^2 - 2\lambda_n \mathcal{R}(\boldsymbol{\theta}^*) \\
&\geq \frac{1}{n}\sum_{t=\hat{t}_l}^{t_k^*-1} \left(\mathbf{x}_t^T\left(\hat{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_{k-1}^*\right)\right)^2 + \frac{1}{n}\sum_{t=t_k^*}^{\hat{t}_{l+1}-1} \left(\mathbf{x}_t^T\left(\hat{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_k^*\right)\right)^2 - 2\lambda_n \mathcal{R}(\boldsymbol{\theta}^*) \\
&> 0,
\end{aligned}$$

where the proof of the last step follows the derivations in Appendix A.9.1 (in particular for the derivation of (A.27)), and it holds with probability at least $1-2c^{'}\exp\left(-cI_{\min}/2\right)$. Following the same reasoning in previous sections, $\hat{K} \geq K^*$ if $1-2c^{'}\exp\left(-cI_{\min}/2\right) > 0$. And from Assumption B1, we have that $1-2c^{'}\exp\left(-cI_{\min}/2\right) > 1-2c^{'}\exp\left(-cn\delta_n\right) > 0$.

A.9.4 Proof for Proposition 10

Suppose there is a true change-point t_k^* such that $|t_k^* - \hat{t}_i| > n\delta_n$ for some $i = 1, \cdots, \hat{K}$ i.e. $\varepsilon\left(\hat{T}_{\hat{K}} || \{t_k^*\}\right) > n\delta_n$. Suppose the nearest estimated change-points are $\hat{t}_l$ and $\hat{t}_{l+1}$. Hence $t_k^* < \hat{t}_{l+1}$ and $t_k^* > \hat{t}_l$ and $\hat{t}_{l+1} - t_k^* > n\delta_n$ and $t_k^* - \hat{t}_l > n\delta_n$. Then following the same technique in Appendix A.9.1, $\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}})$ will be larger than 0 with a probability at least $1 - 2c^{'} \exp\left(-cn\delta_n\right)$. So we have

$$\mathbb{P}\left\{\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) > 0 \,\middle|\, \varepsilon\left(\hat{\mathbb{T}}_{\hat{K}} || \{t_k^*\}\right) > n\delta_n\right\} \geq 1 - 2c^{'} \exp\left(-cn\delta_n\right) > 0, \tag{A.37}$$

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

$$\varepsilon\left(\hat{\mathbb{T}}_{\hat{K}} || \mathbb{T}^*\right) \leq n\delta_n.$$

A.9.5 Proof for Proposition 11

In this proof, we focus on the case illustrated in Fig. 3.1 and other cases follow similarly. Suppose $\mathbb{T}_{isolate}$ happens, i.e., there exists some isolated estimated change-points $\hat{t}_j$ with

$\hat{t}_{j+1} - \hat{t}_j \geq n\delta_n, \hat{t}_{j+1} \leq t^*_{k+1}, \hat{t}_j - \hat{t}_{j-1} \geq n\delta_n, \hat{t}_{j-1} \geq t^*_k, \|\hat{\boldsymbol{\alpha}}_{j+1} - \hat{\boldsymbol{\alpha}}_j\|_2 > J_{\min}$, we have

$$
\begin{aligned}
&\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) \\
= \;& \frac{1}{n}\|\mathbf{X}\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2^2 - 2\lambda_n \mathcal{R}(\boldsymbol{\theta}^*) \\
\geq \;& \frac{1}{n}\sum_{t=\hat{t}_{j-1}}^{\hat{t}_j - 1} \left(\mathbf{x}_t^T \left(\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*_{k+1}\right)\right)^2 + \frac{1}{n}\sum_{t=\hat{t}_j}^{\hat{t}_{j+1}-1} \left(\mathbf{x}_t^T \left(\hat{\boldsymbol{\alpha}}_{j+1} - \boldsymbol{\alpha}^*_{k+1}\right)\right)^2 - 2\lambda_n \mathcal{R}(\boldsymbol{\theta}^*) \\
\overset{(a)}{\geq} \;& \delta_n \frac{l}{32}\left(\|\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*_{k+1}\|_2^2 + \|\hat{\boldsymbol{\alpha}}_{j+1} - \boldsymbol{\alpha}^*_{k+1}\|_2^2\right) - \frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) - 2\lambda_n \mathcal{R}(\boldsymbol{\theta}^*) \\
\geq \;& \delta_n \frac{l}{32}\frac{1}{2}\|\hat{\boldsymbol{\alpha}}_{j+1} - \hat{\boldsymbol{\alpha}}_j\|_2^2 - \frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) - 2\lambda_n \mathcal{R}(\boldsymbol{\theta}^*) \\
> \;& \frac{l}{64}\delta_n J_{\min}^2 - \frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) - 2\lambda_n \mathcal{R}(\boldsymbol{\theta}^*) \\
= \;& 0.
\end{aligned}
$$

where step (a) follows (A.34), (A.35) and (A.36) and the last step follows the definition of δ_n, and this result holds with probability $1 - 2c^{'}\exp\left(-cn\delta_n\right)$. So we have

$$
\mathbb{P}\left\{\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) > 0 \mid \mathbb{T}_{isolate}\right\} \geq 1 - 2c^{'}\exp\left(-cn\delta_n\right) > 0, \tag{A.38}
$$

where the last step is due to Assumption B1. As $\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}})$ must be less or equal than 0, we have that the event $\mathbb{T}_{isolate}$ does not occur.

A.9.6 Proof for Proposition 12

Following the proof in Appendix A.9.1, for $i \in \Upsilon\left(\frac{I_{\min}}{10n}\right)$, we have

$$
\begin{aligned}
&\mathcal{F}(\mathbf{\Delta}_{\boldsymbol{\theta}}) \\
\geq\ & \frac{1}{n}\|\mathbf{X}\mathbf{\Delta}_{\boldsymbol{\beta}}\|_2^2 - 2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*) \\
\geq\ & \frac{1}{n}\sum_{t=\hat{t}_{i-1}}^{\hat{t}_i-1}\left(\mathbf{x}_t\left(\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}^*_{k(i)}\right)\right)^2 - 2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*) && \text{(A.39)} \\
\geq\ & \left(\frac{\hat{t}_i - \hat{t}_{i-1}}{n}\frac{l}{32}\|\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}^*_{k(i)}\|_2^2 - \frac{36}{(1-\gamma)^2}\sqrt{L}\rho(\mathbf{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*)\right) \\
& -2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*) && \text{(A.40)} \\
>\ & \frac{I_{\min}}{10n}\frac{l}{32}\|\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}^*_{k(i)}\|_2^2 - \frac{36}{(1-\gamma)^2}\sqrt{L}\rho(\mathbf{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) \\
& -2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*). && \text{(A.41)}
\end{aligned}
$$

The probability that (A.40) holds is at least $1-c^{'}\exp\left(-cI_{\min}/10\right) > 0$ due to Assumption B1 leading to the fact that $I_{\min} \to \infty$ as $n \to \infty$. (A.41) uses the fact that $i \in \Upsilon\left(\frac{I_{\min}}{10n}\right)$ hence $\hat{t}_i - \hat{t}_{i-1} > \frac{I_{\min}}{10n}$. Following similar steps in Appendix A.9.2, for $i \in \Upsilon\left(\frac{I_{\min}}{10n}\right)$, we have

$$
\|\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}^*_{k(i)}\|_2^2 \leq \frac{320n}{I_{\min}l}\left(\frac{36}{(1-\gamma)^2}\sqrt{L}\rho(\mathbf{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) + 2\lambda_n\mathcal{R}(\boldsymbol{\theta}^*)\right).
$$

A.10 Proof for Simplied Results in Section 3.2.3

A.10.1 Proof of the first and second item in Lemma 3

We first find an upper bound of δ_n. Under the additional assumption A3 and we choose $\lambda_n = \frac{8\sigma\sqrt{5\rho_c}}{1-\gamma}\sqrt{\frac{\log np}{n}}$, using the definition of δ_n in (3.10), we have that

$$\begin{aligned} \delta_n &\overset{(a)}{\leq} c_3\left(s^2\sqrt{\frac{\log p}{n}} + s\sqrt{\frac{\log np}{n}}\right), \\ &\overset{(b)}{\leq} c_4 s^2\sqrt{\frac{\log p}{n}}, \end{aligned} \tag{A.42}$$

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

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

$$\delta_n \geq \frac{64}{l}\left(\frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\left(\frac{\mathcal{R}(\boldsymbol{\theta}^*)}{J_{\min}}\right)^2\right) \geq c_5\sqrt{\frac{\log p}{n}}. \tag{A.43}$$

For the third item in Lemma 3, following the same reasoning for finding bounds for δ_n, we can show that the right hand side of (3.13) can be upper bounded by

$$c_8 s^2\sqrt{\frac{\log p}{n}}. \tag{A.44}$$

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

A.10.2 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 [80, 81].

Lemma 10. 1. *Dual norm of* $\sum_{t=1}^{n} \|\boldsymbol{\theta}_t\|_2$ *is* $\max_{1 \le t \le n} \|\boldsymbol{\theta}_t\|_2$.

2. *Dual norm of* ℓ_1 *is* ℓ_∞.

3. *Let* Ξ_1 *and* Ξ_2 *be two norms, if* $\Xi = \Xi_1 + \Xi_2$, *then*

$$\Xi^*(\boldsymbol{\theta}) = \min_{\boldsymbol{z}} \max\{\Xi_1(\boldsymbol{\theta} - \boldsymbol{z}), \Xi_2(\boldsymbol{z})\}. \tag{A.45}$$

4. *Let* $\alpha \in \mathbb{R}^+$ *be a constant, then*

$$\Xi^*(\alpha\boldsymbol{\theta}) = \Xi^*(\boldsymbol{\theta})/\alpha. \tag{A.46}$$

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

Lemma 11.

$$\mathcal{R}^*(\boldsymbol{\theta}) \le \|\boldsymbol{\theta}\|_\infty/(1-\gamma). \tag{A.47}$$

Proof. Applying (A.45) and (A.46), we have

$$\begin{aligned} \mathcal{R}^*(\boldsymbol{\theta}) &= \min_{\boldsymbol{z} \in \mathbb{R}^{np}} \left\{ \max \left\{ \max_{1 \le t \le n} \|(\boldsymbol{\theta} - \boldsymbol{z})_t\|_2/\gamma, \|\boldsymbol{z}\|_\infty/(1-\gamma) \right\} \right\} \\ &\le \|\boldsymbol{\theta}\|_\infty/(1-\gamma). \end{aligned}$$

□

Now, we state a lemma showing that by choosing $\lambda_n = \frac{8\sigma\sqrt{5\rho_c}}{1-\gamma}\sqrt{\frac{\log np}{n}}$ we can satisfy the left side of (3.11) with a high probability.

Lemma 12. *If A1 and A2 hold, and we choose* $\lambda_n \ge \frac{8\sigma\sqrt{5\rho_c}}{1-\gamma}\sqrt{\frac{\log np}{n}}$, *then*

$$\lambda_n \ge 2\mathcal{R}^*(\nabla\mathcal{L}(\boldsymbol{\theta}^*)). \tag{A.48}$$

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

Proof. First, we have

$$\nabla \mathcal{L}(\boldsymbol{\theta}^*) \;\;=\;\; -\frac{2}{n}\tilde{\mathbf{X}}^T \mathbf{e}. \tag{A.49}$$

Using Lemma 11 we have,

$$2\mathcal{R}^*(\nabla \mathcal{L}(\boldsymbol{\theta}^*)) \leq \frac{4}{n(1-\gamma)} \|\tilde{\mathbf{X}}^T \mathbf{e}\|_\infty. \tag{A.50}$$

To bound $\|\tilde{\mathbf{X}}^T \mathbf{e}\|_\infty$, we first study the first p columns in $\tilde{\mathbf{X}}$ and other column follows. For $1 \leq j \leq p$, we have

$$\begin{aligned}
& \mathbb{P}\left\{\left|\frac{1}{n}\left(\tilde{\mathbf{X}}_{\cdot,j}\right)^T \mathbf{e}\right| \geq c\right\} \\
\leq \;\; & \mathbb{P}\left\{\left|\frac{1}{n}\left(\tilde{\mathbf{X}}_{\cdot,j}\right)^T \mathbf{e}\right| \geq c, \frac{\|\tilde{\mathbf{X}}_{\cdot,j}\|_2}{\sqrt{n}} \leq \sqrt{5\boldsymbol{\Sigma}_{j,j}}\right\} + \mathbb{P}\left\{\left|\frac{1}{n}\left(\tilde{\mathbf{X}}_{\cdot,j}\right)^T \mathbf{e}\right| \geq c, \frac{\|\tilde{\mathbf{X}}_{\cdot,j}\|_2}{\sqrt{n}} > \sqrt{5\boldsymbol{\Sigma}_{j,j}}\right\}.
\end{aligned}$$

Noticing that each element in $\tilde{\mathbf{X}}_{\cdot,j}$ is a realization, we have

$$\mathbb{P}\left\{\left|\frac{1}{n}\left(\tilde{\mathbf{X}}_{\cdot,j}\right)^T \mathbf{e}\right| \geq c, \frac{\|\tilde{\mathbf{X}}_{\cdot,j}\|_2}{\sqrt{n}} \leq \sqrt{5\boldsymbol{\Sigma}_{j,j}}\right\} \leq 2\exp\left(-\frac{nc^2}{10\sigma^2\boldsymbol{\Sigma}_{j,j}}\right), \tag{A.51}$$

which follows from the Gaussian tail bound. In addition, we have

$$\begin{aligned}
& \mathbb{P}\left\{\left|\frac{1}{n}\left(\tilde{\mathbf{X}}_{\cdot,j}\right)^T \mathbf{e}\right| \geq c, \frac{\|\tilde{\mathbf{X}}_{\cdot,j}\|_2}{\sqrt{n}} > \sqrt{5\boldsymbol{\Sigma}_{j,j}}\right\} \\
\leq \;\; & \mathbb{P}\left\{\frac{\|\tilde{\mathbf{X}}_{\cdot,j}\|_2}{\sqrt{n}} > \sqrt{5\boldsymbol{\Sigma}_{j,j}}\right\} \\
= \;\; & \mathbb{P}\left\{\frac{\|\tilde{\mathbf{X}}_{\cdot,j}\|_2^2}{\boldsymbol{\Sigma}_{j,j}} > 5n\right\} \\
\leq \;\; & \exp(-n),
\end{aligned}$$

where the last step is due to the chi-square tail bound. Combining these two, we have

$$\mathbb{P}\left\{\left|\frac{1}{n}\left(\tilde{\mathbf{X}}_{\cdot,j}\right)^T\mathbf{e}\right| \geq c\right\} \;\leq\; 2\exp\left(-\frac{nc^2}{10\sigma^2\mathbf{\Sigma}_{j,j}}\right) + \exp\left(-n\right).$$

As the result, we obtain

$$\begin{aligned}\mathbb{P}\left\{\left\|\frac{1}{n}\tilde{\mathbf{X}}^T\mathbf{e}\right\|_\infty \geq c\right\} \;&\leq\; np\left(2\exp\left(-\frac{nc^2}{10\sigma^2\mathbf{\Sigma}_{j,j}}\right) + \exp\left(-n\right)\right)\\ &=\; 2\exp\left(-\frac{nc^2}{10\sigma^2\mathbf{\Sigma}_{j,j}} + \log np\right) + \exp\left(-n + \log np\right),\end{aligned}$$

which leads to

$$\mathbb{P}\left\{\left\|\frac{4/(1-\gamma)}{n}\tilde{\mathbf{X}}^T\mathbf{e}\right\|_\infty \geq c\right\} \;\leq\; 2\exp\left(-\frac{nc^2}{160\sigma^2\mathbf{\Sigma}_{j,j}/(1-\gamma)^2} + \log np\right) + \exp\left(-n + \log np\right).$$

Setting $c^2 = \frac{320/(1-\gamma)^2\sigma^2\rho_c\log np}{n}$ and $\lambda_n \geq \frac{8\sigma\sqrt{5\rho_c}}{1-\gamma}\sqrt{\frac{\log np}{n}}$, we obtain $\mathbb{P}\left\{\left\|\frac{4/(1-\gamma)}{n}\tilde{\mathbf{X}}^T\mathbf{e}\right\|_\infty < \lambda_n\right\} \geq 1 - \frac{2}{np} - np\exp\left(-n\right)$. □

Now, we show our choice of λ_n satisfies the right side of (3.11). Since $s^2\sqrt{\frac{\log p}{n}} \to 0$ in A3, then using (A.42), we know that $\delta_n \to 0$. Combining this with the assumption $I_{\min} = \Theta(n)$ in A3, we have

$$\frac{I_{\min}}{2n} \geq c_6 > \delta_n \to 0 \text{ as } n \to \infty,$$

and

$$\frac{I_{\min}}{2n} - \delta_n \geq c_7. \tag{A.52}$$

This means we can satisfy the condition $\frac{I_{\min}}{2n} > \delta_n$, i.e.,

$$\lambda_n \leq \frac{1}{2\mathcal{R}(\boldsymbol{\theta}^*)}\left(\frac{I_{\min}}{2n}\frac{lJ_{\min}^2}{64} - \frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\mathbf{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*)\right), \tag{A.53}$$

which is the upper bound in (3.11).

Combing the result of Lemma 12 and (A.53), we have finished the proof for the third item in Lemma 3.

A.11 Proof for results on GLM in Section 3.3

Here we follow the same proof outline as the Appendix A.9, and only show the key steps.

A.11.1 Supporting Results

We first provide several supporting results that will be used throughout the proof.

First, as we choose

$$\lambda_n \geq 2\mathcal{R}^*(\nabla\mathcal{L}(\boldsymbol{\theta}^*)), \tag{A.54}$$

then following similar steps as those in the proof of Lemma 1, we can prove

$$\mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) \quad \leq \quad 4\mathcal{R}(\boldsymbol{\theta}^*). \tag{A.55}$$

Second, motivated the proofs in Section 3.2.2 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. Again we have $\hat{\mathbb{T}}_{\hat{K}}$ and $\mathbb{T}^*$ whose definitions are the same as those in Section 3.2.2. Let $\left|\mathbb{T}^* \cup \hat{\mathbb{T}}_{\hat{K}}\right| = \tilde{K}$ and $\mathbb{T}^* \cup \hat{\mathbb{T}}_{\hat{K}} = \{\tilde{t}_i\}_{i=1}^{\tilde{K}}$. In the following, we first show a property related to two adjacent intervals $[\tilde{t}_{j-1}, \tilde{t}_j) \cup [\tilde{t}_j, \tilde{t}_{j+1}) = [\tilde{t}_{j-1}, \tilde{t}_{j+1})$.

We have

$$\begin{aligned}
\delta\mathcal{L}(\mathbf{\Delta}_{\boldsymbol{\theta}},\boldsymbol{\theta}^*): &= \mathcal{L}(\boldsymbol{\theta}^*+\mathbf{\Delta}_{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^*) - \langle \nabla\mathcal{L}(\boldsymbol{\theta}^*), \mathbf{\Delta}_{\boldsymbol{\theta}}\rangle \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(-y_t\langle \mathbf{\Delta}_{\boldsymbol{\theta}}, \tilde{\mathrm{x}}_t\rangle + \Phi\left(\langle \boldsymbol{\theta}^*+\mathbf{\Delta}_{\boldsymbol{\theta}}, \tilde{\mathrm{x}}_t\rangle\right) - \Phi\left(\langle \boldsymbol{\theta}^*, \tilde{\mathrm{x}}_t\rangle\right)\right) \\
&\quad -\left\langle \frac{1}{n}\sum_{t=1}^{n}\left(-y_t\tilde{\mathrm{x}}_t + \tilde{\mathrm{x}}_t\Phi^{'}\left(\langle \boldsymbol{\theta}^*, \tilde{\mathrm{x}}_t\rangle\right)\right), \mathbf{\Delta}_{\boldsymbol{\theta}}\right\rangle \\
&= \frac{1}{n}\sum_{t=1}^{n}\Phi\left(\langle \boldsymbol{\theta}^*+\mathbf{\Delta}_{\boldsymbol{\theta}}, \tilde{\mathrm{x}}_t\rangle\right) - \frac{1}{n}\sum_{t=1}^{n}\Phi\left(\langle \boldsymbol{\theta}^*, \tilde{\mathrm{x}}_t\rangle\right) - \frac{1}{n}\sum_{t=1}^{n}\langle \tilde{\mathrm{x}}_t\Phi^{'}\left(\langle \boldsymbol{\theta}^*, \tilde{\mathrm{x}}_t\rangle\right), \mathbf{\Delta}_{\boldsymbol{\theta}}\rangle.
\end{aligned}$$

If Φ is a strictly convex function, then $\Phi^{''} > 0$. Using Taylor expansion, we know that there exists $\boldsymbol{\xi}_t \in \mathbb{R}^{np}$ such that

$$\begin{aligned}
\delta\mathcal{L}(\mathbf{\Delta}_{\boldsymbol{\theta}},\boldsymbol{\theta}^*) &= \mathbf{\Delta}_{\boldsymbol{\theta}}^T\left(\frac{1}{n}\sum_{t=1}^{n}\tilde{\mathrm{x}}_t\tilde{\mathrm{x}}_t^T\Phi^{''}(\langle \boldsymbol{\xi}_t, \tilde{\mathrm{x}}_t\rangle)\right)\mathbf{\Delta}_{\boldsymbol{\theta}} \\
&\geq \frac{\mu}{n}\|\tilde{\mathbf{X}}\mathbf{\Delta}_{\boldsymbol{\theta}}\|_2^2 \\
&= \frac{\mu}{n}\|\mathbf{X}\mathbf{\Delta}_{\boldsymbol{\beta}}\|_2^2,
\end{aligned}$$

where $\mu = \min_{1\leq t\leq n}\{\Phi^{''}(\langle \boldsymbol{\xi}_t, \tilde{\mathrm{x}}_t\rangle)\} > 0$ and its value depends on the function Φ, data matrix $\mathbf{X}$ and true parameter $\boldsymbol{\theta}^*$.

Now, we focus on two adjacent intervals $[\tilde{t}_{j-1}, \tilde{t}_j) \cup [\tilde{t}_j, \tilde{t}_{j+1}) = [\tilde{t}_{j-1}, \tilde{t}_{j+1})$. Since $(\mathbf{\Delta}_{\boldsymbol{\beta}})_t$ is the same for $t \in [\tilde{t}_{j-1}, \tilde{t}_j)$ for $j \in [\tilde{K}+1]$, then let $(\tilde{\mathbf{\Delta}}_{\boldsymbol{\beta}})_j$ denote the coefficients difference vector of the interval $t \in [\tilde{t}_{j-1}, \tilde{t}_j)$:

$$
\begin{aligned}
& \frac{\mu}{n}\|\mathbf{X}\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2^2 \\
\geq\ & \sum_{i=j}^{j+1} \frac{\mu}{n} \left\| \begin{pmatrix} \mathbf{x}_{\tilde{t}_{i-1}} \\ \cdots \\ \mathbf{x}_{\tilde{t}_i - 1} \end{pmatrix} (\tilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})_i \right\|_2^2 \\
\overset{(a)}{\geq}\ & \sum_{i=j}^{j+1} \mu \frac{\tilde{t}_i - \tilde{t}_{i-1}}{n} \left(\frac{l}{32}\|(\tilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})_i\|_2^2 - \frac{9}{4}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{\tilde{t}_i - \tilde{t}_{i-1}}}\|(\tilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})_i\|_2\|(\tilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})_i\|_1 \right) \\
\overset{(b)}{\geq}\ & \frac{\mu l}{32}\left(\frac{\tilde{t}_j - \tilde{t}_{j-1}}{n}\|(\tilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})_j\|_2^2 + \frac{\tilde{t}_{j+1} - \tilde{t}_j}{n}\|(\tilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})_{j+1}\|_2^2 \right) - \mu\frac{9}{4}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\sum_{i=j}^{j+1}\|(\tilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})_i\|_2\|(\tilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})_i\|_1 \\
\overset{(c)}{\geq}\ & \frac{\mu l}{32}\left(\frac{\tilde{t}_j - \tilde{t}_{j-1}}{n}\|(\tilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})_j\|_2^2 + \frac{\tilde{t}_{j+1} - \tilde{t}_j}{n}\|(\tilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})_{j+1}\|_2^2 \right) - \mu\frac{9}{2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\|\boldsymbol{\Delta}_{\boldsymbol{\theta}}\|_1^2, \\
\overset{(d)}{\geq}\ & \frac{\mu l}{32}\left(\frac{\tilde{t}_j - \tilde{t}_{j-1}}{n}\|(\tilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})_j\|_2^2 + \frac{\tilde{t}_{j+1} - \tilde{t}_j}{n}\|(\tilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})_{j+1}\|_2^2 \right) - \mu\frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*), \quad \text{(A.56)}
\end{aligned}
$$

where (a) uses Lemma 2, in (b) we use the fact $\tilde{t}_j - \tilde{t}_{j-1} \leq n$, (c) uses the fact that $\|(\tilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})_i\|_2 \leq \|(\tilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})_i\|_1 \leq \|\boldsymbol{\Delta}_{\boldsymbol{\theta}}\|_1$, (d) uses $\|\boldsymbol{\Delta}_{\boldsymbol{\theta}}\|_1 \leq \frac{1}{1-\gamma}\mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) \leq \frac{4}{1-\gamma}\mathcal{R}(\boldsymbol{\theta}^*)$, as shown in (A.55).

Next, we will use property (A.56) to study $\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}})$.

$$
\begin{aligned}
\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) &= \mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^*) + \lambda_n\left(\mathcal{R}(\hat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^*)\right) \\
&= \delta\mathcal{L}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) + \langle \nabla\mathcal{L}(\boldsymbol{\theta}^*), \boldsymbol{\Delta}_{\boldsymbol{\theta}}\rangle + +\lambda_n\left(\mathcal{R}(\hat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^*)\right) \\
&\overset{(a)}{\geq} \delta\mathcal{L}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \frac{\lambda_n}{2}\mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) - \lambda_n\mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) \\
&\overset{(b)}{\geq} \delta\mathcal{L}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - 6\lambda_n\mathcal{R}(\boldsymbol{\theta}^*) \\
&\overset{(c)}{\geq} \frac{\mu l}{32}\left(\frac{\tilde{t}_j - \tilde{t}_{j-1}}{n}\|(\tilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})_j\|_2^2 + \frac{\tilde{t}_{j+1} - \tilde{t}_j}{n}\|(\tilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})_{j+1}\|_2^2 \right) \\
&\quad -\mu\frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) - 6\lambda_n\mathcal{R}(\boldsymbol{\theta}^*). \qquad \text{(A.57)}
\end{aligned}
$$

where (a) uses the fact that we choose $\lambda_n \geq 2\mathcal{R}^*(\nabla\mathcal{L}(\boldsymbol{\theta}^*))$ and triangle inequality, (b) uses $\mathcal{R}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) \leq 4\mathcal{R}(\boldsymbol{\theta}^*)$ from (A.55) and (c) uses (A.56).

In the following proof, we will use (A.57) repeatedly. In the following, we only provide the outline of the proofs as other details involving probabilities are similar to those in Appendix A.9.

A.11.2 Proof for Proposition 13

The core of the proof is similar to Appendix A.9.1. Here, we focus on only one case as all other cases are similar.

Suppose $t_j^* - \hat{t}_j > n\delta_n$, $n > \hat{t}_{j+1} - t_j^* > \frac{I_{\min}}{2}$ and $\hat{t}_{j+1} \leq t_{j+1}^*$, we have

$$\begin{aligned}
\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) &\geq \frac{\mu l}{32}\left(\frac{t_j^* - \tilde{t}_j}{n}\|\boldsymbol{\alpha}_j^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_2^2 + \frac{\hat{t}_{j+1} - t_j^*}{n}\|\boldsymbol{\alpha}_{j+1}^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_2^2\right) \\
&\quad -\mu\frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) - 6\lambda_n\mathcal{R}(\boldsymbol{\theta}^*) \\
&> \frac{\mu l}{32}\left(\delta_n\|\boldsymbol{\alpha}_j^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_2^2 + \frac{I_{\min}}{2n}\|\boldsymbol{\alpha}_{j+1}^* - \hat{\boldsymbol{\alpha}}_{j+1}\|_2^2\right) \\
&\quad -\mu\frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) - 6\lambda_n\mathcal{R}(\boldsymbol{\theta}^*) \\
&\geq \frac{\mu l}{64}\delta_n\|\boldsymbol{\alpha}_j^* - \boldsymbol{\alpha}_{j+1}^*\|_2^2 - \mu\frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) - 6\lambda_n\mathcal{R}(\boldsymbol{\theta}^*) \\
&\geq \frac{\mu l}{64}\delta_n J_{\min}^2 - \mu\frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\boldsymbol{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) - 6\lambda_n\mathcal{R}(\boldsymbol{\theta}^*) \\
&= 0, \qquad\qquad (A.58)
\end{aligned}$$

which is an contradiction. Hence the conclusion in Proposition 13 holds.

A.11.3 Proof for Proposition 14

Since Proposition 13 holds, we have $|\hat{t}_k - t_k^*| \leq n\delta_n$ for each $k = 1, \cdots, K^*$. For each $k \in [K^*]$, we have

$$\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) \quad \geq \quad \frac{\mu l}{32} \frac{I_{\min} - 2n\delta_n}{n} \|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\|_2^2 - \mu \frac{72}{(1-\gamma)^2} \sqrt{L} \rho(\boldsymbol{\Sigma}) \sqrt{\frac{\log p}{n}} \mathcal{R}^2(\boldsymbol{\theta}^*) - 6\lambda_n \mathcal{R}(\boldsymbol{\theta}^*).$$

Hence, for $\mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) \leq 0$ to hold, we have

$$\|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\|_2^2 \leq \frac{32n}{\mu l (I_{\min} - 2n\delta_n)} \left(\mu \frac{72}{(1-\gamma)^2} \sqrt{L} \rho(\boldsymbol{\Sigma}) \sqrt{\frac{\log p}{n}} \mathcal{R}^2(\boldsymbol{\theta}^*) + 6\lambda_n \mathcal{R}(\boldsymbol{\theta}^*) \right). \tag{A.59}$$

A.11.4 Proof for Proposition 15

Suppose $\hat{K} < K^*$, then there exists at least one true change-point t_k^* such that there is no estimated change-points in the range of $I_{\min}/2$ around it, which is illustrated in Figure A.4. We have

$$\begin{aligned} \mathcal{F}(\boldsymbol{\Delta}_{\boldsymbol{\theta}}) \quad &\geq \quad \frac{\mu l}{32} \left(\frac{t_k^* - \hat{t}_l}{n} \|\hat{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_{k-1}^*\|_2^2 + \frac{\hat{t}_{l+1} - t_k^*}{n} \|\hat{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_k^*\|_2^2 \right) \\ &\qquad -\mu \frac{72}{(1-\gamma)^2} \sqrt{L} \rho(\boldsymbol{\Sigma}) \sqrt{\frac{\log p}{n}} \mathcal{R}^2(\boldsymbol{\theta}^*) - 6\lambda_n \mathcal{R}(\boldsymbol{\theta}^*) \\ &\geq \quad \frac{\mu l}{64} \frac{I_{\min}}{2n} J_{\min}^2 - \mu \frac{72}{(1-\gamma)^2} \sqrt{L} \rho(\boldsymbol{\Sigma}) \sqrt{\frac{\log p}{n}} \mathcal{R}^2(\boldsymbol{\theta}^*) - 6\lambda_n \mathcal{R}(\boldsymbol{\theta}^*) \\ &> \quad 0, \end{aligned}$$

which is a contradiction. Hence $\hat{K} \geq K^*$.

A.11.5 Proof for Proposition 16

Suppose there is some true change-point t_k^* such that $|t_k^* - \hat{t}_i| > n\delta_n$ for $i = 1, \cdots, \hat{K}$. Suppose the nearest estimated change-points are $\hat{t}_l$ and $\hat{t}_{l+1}$. Hence $t_k^* < \hat{t}_{l+1}$ and $t_k^* > \hat{t}_l$ and $\hat{t}_{l+1} - t_k^* > n\delta_n$ and $t_k^* - \hat{t}_l > n\delta_n$. Then following the same technique in Appendix A.11.2 we can find that

$$\mathcal{F}(\mathbf{\Delta}_{\boldsymbol{\theta}}) > 0,$$

which is a contradiction. Hence $\frac{\epsilon\left(\hat{\mathbb{T}}_{\hat{K}} \| \mathbb{T}^*\right)}{n} \leq \delta_n$.

A.11.6 Proof for Proposition 17

In this proof, we focus on the case illustrated in Figure 3.1 and other cases follow similarly. Suppose there exists some isolated estimated change-points $\hat{t}_j$ with $\hat{t}_{j+1} - \hat{t}_j \geq n\delta_n$, $\hat{t}_{j+1} \leq t_{k+1}^*$, $\hat{t}_j - \hat{t}_{j-1} \geq n\delta_n$, $\hat{t}_{j-1} \geq t_k^*$, $\|\hat{\boldsymbol{\alpha}}_{j+1} - \hat{\boldsymbol{\alpha}}_j\|_2 > J_{\min}$, we have

$$\begin{aligned}
\mathcal{F}(\mathbf{\Delta}_{\boldsymbol{\theta}}) \;&\geq\; \frac{\mu l}{32}\left(\frac{\hat{t}_{j+1} - \hat{t}_j}{n}\|\hat{\boldsymbol{\alpha}}_{j+1} - \boldsymbol{\alpha}_{k+1}^*\|_2^2 + \frac{\hat{t}_j - \hat{t}_{j-1}}{n}\|\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_{k+1}^*\|_2^2\right) \\
&\qquad - \mu\frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\mathbf{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) - 6\lambda_n\mathcal{R}(\boldsymbol{\theta}^*) \\
&>\; \frac{\mu l}{64}\delta_n J_{\min}^2 - \mu\frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\mathbf{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) - 6\lambda_n\mathcal{R}(\boldsymbol{\theta}^*) \\
&=\; 0.
\end{aligned}$$

This means that the event $\mathbb{T}_{isolate}$ cannot happen.

A.11.7 Proof for Proposition 18

Following the steps in previous appendices, we have

$$\mathcal{F}(\boldsymbol{\theta}) \;\geq\; \frac{\mu l}{32}\frac{I_{\min}}{10n}\|\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_{k(i)}^*\|_2^2 - \mu\frac{72}{(1-\gamma)^2}\sqrt{L}\rho(\mathbf{\Sigma})\sqrt{\frac{\log p}{n}}\mathcal{R}^2(\boldsymbol{\theta}^*) - 6\lambda_n\mathcal{R}(\boldsymbol{\theta}^*).$$

Hence for $\mathcal{F}(\boldsymbol{\theta}) \leq 0$ to be true, we must have

$$\|\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}^*_{k(i)}\|_2^2 \leq \frac{320n}{\mu l I_{\min}} \left(\mu \frac{72}{(1-\gamma)^2} \sqrt{L} \rho(\boldsymbol{\Sigma}) \sqrt{\frac{\log p}{n}} \mathcal{R}^2(\boldsymbol{\theta}^*) + 6\lambda_n \mathcal{R}(\boldsymbol{\theta}^*) \right).$$

A.12 Proof of Proposition 19

For the first item in the proposition,

$$\begin{aligned}
\varphi(\boldsymbol{\beta}_{k+1}) &= \mathcal{L}(\boldsymbol{\beta}_{k+1}) + \mathcal{R}(\boldsymbol{\beta}_{k+1}) \\
&\leq \mathcal{L}(\boldsymbol{\beta}_k) + \langle \nabla \mathcal{L}(\boldsymbol{\beta}_k), \boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k \rangle + \frac{L}{2} \|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|_2^2 + \mathcal{R}(\boldsymbol{\beta}_{k+1}) \\
&\overset{(a)}{\leq} \mathcal{L}(\boldsymbol{\beta}_k) + \langle \nabla \mathcal{L}(\boldsymbol{\beta}_k), \boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k \rangle + \frac{L}{2} \|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|_2^2 \\
&\quad + \mathcal{R}(\boldsymbol{\beta}_k) + \langle \mathrm{s}(\boldsymbol{\beta}_{k+1}), \boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k \rangle \\
&= \varphi(\boldsymbol{\beta}_k) + \langle \nabla \mathcal{L}(\boldsymbol{\beta}_k) + \mathrm{s}(\boldsymbol{\beta}_{k+1}), \boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k \rangle + \frac{L}{2} \|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|_2^2 \\
&\overset{(b)}{=} \varphi(\boldsymbol{\beta}_k) + \langle \nabla \mathcal{L}(\boldsymbol{\beta}_k) - \nabla \mathcal{L}(\boldsymbol{\beta}_{fixed}), \boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k \rangle - \frac{L}{2} \|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|_2^2 \\
&\leq \varphi(\boldsymbol{\beta}_k) + L \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{fixed}\|_2 \|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|_2 - \frac{L}{2} \|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|_2^2 \\
&\leq \varphi(\boldsymbol{\beta}_k). \qquad \text{(A.60)}
\end{aligned}$$

where (a) uses (4.3), (b) uses $\nabla \mathcal{L}(\boldsymbol{\beta}_{fixed}) + L(\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k) + \mathrm{s}(\boldsymbol{\beta}_{k+1}) = \mathbf{0}$.

Then we prove the second item.

$$
\begin{aligned}
\varphi(\boldsymbol{\beta}_{k+1}) &= \mathcal{L}(\boldsymbol{\beta}_{k+1}) + \mathcal{R}(\boldsymbol{\beta}_{k+1}) \\
&\leq \mathcal{L}(\boldsymbol{\beta}_k) + \langle \nabla \mathcal{L}(\boldsymbol{\beta}_k), \boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k \rangle + \frac{L}{2}\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|_2^2 + \mathcal{R}(\boldsymbol{\beta}_{k+1}) \\
&\overset{(c)}{\leq} \mathcal{L}(\boldsymbol{\beta}^*) + \langle \nabla \mathcal{L}(\boldsymbol{\beta}_k), \boldsymbol{\beta}_k - \boldsymbol{\beta}^* \rangle - \frac{\mu_{\mathcal{L}}}{2}\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|_2^2 \\
&\quad + \langle \nabla \mathcal{L}(\boldsymbol{\beta}_k), \boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k \rangle + \frac{L}{2}\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|_2^2 \\
&\quad + \mathcal{R}(\boldsymbol{\beta}^*) + \langle \mathrm{s}(\boldsymbol{\beta}_{k+1}), \boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}^* \rangle \\
&= \varphi(\boldsymbol{\beta}^*) + \langle \nabla \mathcal{L}(\boldsymbol{\beta}_k) + \mathrm{s}(\boldsymbol{\beta}_{k+1}), \boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}^* \rangle \\
&\quad + \frac{L}{2}\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|_2^2 - \frac{\mu_{\mathcal{L}}}{2}\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|_2^2 \\
&\overset{(d)}{=} \varphi(\boldsymbol{\beta}^*) + \langle \nabla \mathcal{L}(\boldsymbol{\beta}_k) - \nabla \mathcal{L}(\boldsymbol{\beta}_{fixed}), \boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}^* \rangle + \langle -L(\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k), \boldsymbol{\beta}_k - \boldsymbol{\beta}^* \rangle \\
&\quad - \frac{L}{2}\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|_2^2 - \frac{\mu_{\mathcal{L}}}{2}\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|_2^2. \qquad \text{(A.61)}
\end{aligned}
$$

where (c) uses Equation (4.2) and (4.3), (d) uses $\nabla \mathcal{L}(\boldsymbol{\beta}_{fixed}) + L(\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k) + \mathrm{s}(\boldsymbol{\beta}_{k+1}) = \mathbf{0}$.

Using the inequality above, we have

$$
\begin{aligned}
&\frac{L}{2}\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}^*\|^2 \\
= \; &\frac{L}{2}\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|^2 + \langle L(\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k), \boldsymbol{\beta}_k - \boldsymbol{\beta}^* \rangle + \frac{L}{2}\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|^2 \\
\overset{\text{(A.61)}}{\leq} \; &\frac{L}{2}\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|^2 + \varphi(\boldsymbol{\beta}^*) - \varphi(\boldsymbol{\beta}_{k+1}) + \langle \nabla \mathcal{L}(\boldsymbol{\beta}_k) - \nabla \mathcal{L}(\boldsymbol{\beta}_{fixed}), \boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}^* \rangle \\
&- \frac{\mu_{\mathcal{L}}}{2}\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|_2^2
\end{aligned}
$$

Let $c = \frac{\mu_{\mathcal{L}}}{5}$, we have

$$\begin{aligned}
\varphi(\boldsymbol{\beta}_{k+1}) + \frac{L}{2}\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}^*\|^2 &\leq \varphi(\boldsymbol{\beta}^*) + \frac{L-\mu_{\mathcal{L}}}{2}\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|^2 + \langle \nabla\mathcal{L}(\boldsymbol{\beta}_k) - \nabla\mathcal{L}(\boldsymbol{\beta}_{fixed}), \boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}^* \rangle \\
&\leq \varphi(\boldsymbol{\beta}^*) + \frac{L-\mu_{\mathcal{L}}}{2}\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|^2 + L\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{fixed}\|\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}^*\| \\
&\leq \varphi(\boldsymbol{\beta}^*) + \frac{L-\mu_{\mathcal{L}}}{2}\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|^2 + c\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}^*\| \\
&\leq \varphi(\boldsymbol{\beta}^*) + \frac{L-\mu_{\mathcal{L}}}{2}\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|^2 + \frac{c}{2}\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|^2 + \frac{3c}{2}\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}^*\|^2.
\end{aligned}$$

By rearranging the terms, we have

$$\begin{aligned}
&\varphi(\boldsymbol{\beta}_{k+1}) - \varphi(\boldsymbol{\beta}^*) + \frac{1}{2}\left(L - \frac{3}{5}\mu_{\mathcal{L}}\right)\left\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}^*\right\|^2 \\
\leq\ &\left(\frac{5L - 4\mu_{\mathcal{L}}}{5L - 3\mu_{\mathcal{L}}}\right)\left(\frac{1}{2}\left(L - \frac{3}{5}\mu_{\mathcal{L}}\right)\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|^2\right).
\end{aligned}$$

A.13 Proof of Proposition 20

Proof. We prove this proposition by induction.

For $k = 0$, the proposition holds trivially.

Now we assume (4.11) holds for $k - 1$.

If we do exact update, from [60, Proposition 4], we have

$$\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\| \leq \left(1 - \frac{\mu_{\mathcal{L}}}{L}\right)\|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*\| \leq \left(1 - \frac{\mu_{\mathcal{L}}}{L}\right)\left(1 - \frac{\mu_{\mathcal{L}} - \mu_0}{L}\right)^{k-1} D \leq \left(1 - \frac{\mu_{\mathcal{L}} - \mu_0}{L}\right)^{k} D.$$

If we do inexact update,

$$\begin{aligned}
\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\| &\leq \left(1 - \frac{\mu_{\mathcal{L}}}{L}\right) \|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*\| + \frac{\|\mathbf{e}_{k-1}\|}{L} \\
&= \left(1 - \frac{\mu_{\mathcal{L}}}{L}\right) \|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*\| + \frac{\|\nabla \mathcal{L}(\boldsymbol{\beta}_{k-1}) - \nabla \mathcal{L}(\boldsymbol{\beta}_{fixed})\|}{L} \\
&\leq \left(1 - \frac{\mu_{\mathcal{L}}}{L}\right) \|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*\| + \|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}_{fixed}\| \\
&\leq \left(1 - \frac{\mu_{\mathcal{L}}}{L}\right) \left(1 - \frac{\mu_{\mathcal{L}} - \mu_0}{L}\right)^{k-1} D + \frac{\mu_0}{L} \left(1 - \frac{\mu_{\mathcal{L}} - \mu_0}{L}\right)^{k-1} D \\
&= \left(1 - \frac{\mu_{\mathcal{L}} - \mu_0}{L}\right)^{k} D.
\end{aligned}$$

□

A.14 Proof of Proposition 21

First, we show convergence rates of exact iterations.

Lemma 13. *If we use exact update,*

$$\begin{aligned}
\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|_2 &\leq \left(1 - \frac{\mu_{\mathcal{L}}}{L}\right) \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\|_2, \\
\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}^*\|_2 &\leq \left(1 - \frac{\mu_{\mathcal{L}}}{L}\right) \|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|_2.
\end{aligned}$$

Proof.

$$\begin{aligned}
\|\boldsymbol{\beta}_{k+1}-\boldsymbol{\beta}_k\|_2^2 &= \left\| prox_{\frac{1}{L}\mathcal{R}}\left(\boldsymbol{\beta}_k - \frac{1}{L}\nabla\mathcal{L}(\boldsymbol{\beta}_k)\right) - prox_{\frac{1}{L}\mathcal{R}}\left(\boldsymbol{\beta}_{k-1} - \frac{1}{L}\nabla\mathcal{L}(\boldsymbol{\beta}_{k-1})\right)\right\|_2^2 \\
&\overset{(a)}{\leq} \|\boldsymbol{\beta}_k - \frac{1}{L}\nabla\mathcal{L}(\boldsymbol{\beta}_k) - \boldsymbol{\beta}_{k-1} + \frac{1}{L}\nabla\mathcal{L}(\boldsymbol{\beta}_{k-1})\|_2^2 \\
&= \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\|_2^2 + \frac{1}{L^2}\|\nabla\mathcal{L}(\boldsymbol{\beta}_k) - \nabla\mathcal{L}(\boldsymbol{\beta}_{k-1})\|_2^2 \\
&\quad -\frac{2}{L}\langle\nabla\mathcal{L}(\boldsymbol{\beta}_k) - \nabla\mathcal{L}(\boldsymbol{\beta}_{k-1}), \boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\rangle \\
&\overset{(b)}{\leq} \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\|_2^2 + \frac{1}{L^2}\|\nabla\mathcal{L}(\boldsymbol{\beta}_k) - \nabla\mathcal{L}(\boldsymbol{\beta}_{k-1})\|_2^2 \\
&\quad -\frac{2}{L}\left(\frac{1}{L+\mu_{\mathcal{L}}}\|\nabla\mathcal{L}(\boldsymbol{\beta}_k) - \nabla\mathcal{L}(\boldsymbol{\beta}_{k-1})\|_2^2 + \frac{L\mu_{\mathcal{L}}}{L+\mu_{\mathcal{L}}}\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\|_2^2\right) \\
&= \frac{L-\mu_{\mathcal{L}}}{L+\mu_{\mathcal{L}}}\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\|_2^2 - \frac{1}{L^2}\frac{L-\mu_{\mathcal{L}}}{L+\mu_{\mathcal{L}}}\|\nabla\mathcal{L}(\boldsymbol{\beta}_k) - \nabla\mathcal{L}(\boldsymbol{\beta}_{k-1})\|_2^2 \\
&\overset{(c)}{\leq} \frac{L-\mu_{\mathcal{L}}}{L+\mu_{\mathcal{L}}}\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\|_2^2 - \frac{\mu_{\mathcal{L}}^2}{L^2}\frac{L-\mu_{\mathcal{L}}}{L+\mu_{\mathcal{L}}}\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\|_2^2 \\
&= \left(1-\frac{\mu_{\mathcal{L}}}{L}\right)^2\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\|_2^2.
\end{aligned}$$

where (a) uses non-expansiveness of proximal operator, (b) uses Theorem 2.1.12 of [79], (c) uses stronly convexity of $\mathcal{L}$. The second inequality can be proved similarly. □

In order to count the number of rounds of communications in the optimization process, we should count the number of iterations needed between inexact updates. Suppose kth iteration we are out of inexact updates loop. Suppose $\boldsymbol{\beta}_{fixed} = \boldsymbol{\beta}_{k-(p+1)}$. We use t to denote the steps needed to satisfy the constraint in Proposition 20. We have $\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\| \leq \|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}_{k-2}\| \leq \cdots \leq \|\boldsymbol{\beta}_{k-p} - \boldsymbol{\beta}_{fixed}\| \leq \frac{\mu_0}{L}\left(1-\frac{\mu_{\mathcal{L}}-\mu_0}{L}\right)^{k-p} D$. Let $q := \frac{1-\mu_{\mathcal{L}}/L}{1-(\mu_{\mathcal{L}}-\mu_0)/L}$. The number of steps of t to enter the next inexact update phase is the t

such that it at least satisfies

$$\begin{aligned}\left(1-\frac{\mu_{\mathcal{L}}}{L}\right)^t \frac{\mu_0}{L}\left(1-\frac{\mu_{\mathcal{L}}-\mu_0}{L}\right)^{k-p} D &\leq \frac{\mu_0}{L}\left(1-\frac{\mu_{\mathcal{L}}-\mu_0}{L}\right)^{k+t} D,\\ q^t &\leq \left(1-\frac{\mu_{\mathcal{L}}-\mu_0}{L}\right)^{p},\\ t &\geq p\frac{\log\left(1-\frac{\mu_{\mathcal{L}}-\mu_0}{L}\right)}{\log q}.\end{aligned}$$

So for inexact phase of p steps, we need an following exact phase of $p\frac{\log\left(1-\frac{\mu_{\mathcal{L}}-\mu_0}{L}\right)}{\log q}$ steps to reenter the inexact phase. Hence we total number of communication round is

$$N\frac{p\frac{\log\left(1-\frac{\mu_{\mathcal{L}}-\mu_0}{L}\right)}{\log q}}{p+p\frac{\log\left(1-\frac{\mu_{\mathcal{L}}-\mu_0}{L}\right)}{\log q}} = N\frac{\log\left(1-\frac{\mu_{\mathcal{L}}-\mu_0}{L}\right)}{\log\left(1-\frac{\mu_{\mathcal{L}}}{L}\right)}.$$

Bibliography

[1] "Uci machine learning repository." `https://archive.ics.uci.edu/ml/datasets/Arcene`.

[2] "Libsvm data: Classification (binary class)." `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`.

[3] "Communities and crime data set." `https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized`.

[4] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society*, vol. 1, pp. 267–288, Jan. 1996.

[5] S.-I. Lee, H. Lee, P. Abbeel, and A. Y. Ng, "Efficient ℓ_1 regularized logistic regression," in *The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, (Boston, MA), Jan. 2006.

[6] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–451, Apr. 2004.

[7] P. Zhao and B. Yu, "On model selection consistency of Lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, Dec. 2006.

[8] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*, vol. 27, pp. 538–557, Nov. 2012.

[9] P. Perron, *Dealing with strucutral breaks*, vol. 1 of *Palgrave Handbook of Econometrics*. New York: Palgrave Macmillan, 2006.

[10] N. Zhang and D. Siegmund, "A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data," *Biometrics*, vol. 63, pp. 22–32, Mar. 2007.

[11] A. J. Gibberd and J. D. B. Nelson, "High dimensional changepoint detection with a dynamic graphical Lasso," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, (Florence, Italy), pp. 2703–2707, May. 2014.

[12] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes - Theory and Application*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[13] H. V. Poor and O. Hadjiliadis, *Quickest Detection*. Cambridge, UK: Cambridge University Press, 2008.

[14] E. C. Hall and R. M. Willett, “Online optimization in dynamic environments,” 2013. Submitted, available at http://arxiv.org/abs/1307.5944.

[15] V. V. Veeravalli and T. Banerjee, “Quickest change detection.” Elsevier: E-reference Signal Processing, 2013. available at [arXiv:1210.5552].

[16] Y. Xie, J. Huang, and R. Willett, “Change-point detection for high-dimensional time series with missing data,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, pp. 12–27, Feb. 2013.

[17] E. Page, “Estimating the point of change in a continuous process,” *Biometrika*, vol. 44, pp. 248–252, 1957.

[18] P. K. Bhattacharya, “Some aspects of change-point analysis,” *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, vol. 23, p. 2854, 1994.

[19] R. Bellman and R. Roth, “Curve fitting by segmented straight lines,” *Journal of the American Statistical Association*, vol. 64, pp. 1079–1084, 1969.

[20] J. Bai, “Estimation of a change point in multiple regression models,” *The Review of Economics and Statistics*, vol. LXXIX, pp. 551–563, 1997.

[21] J. Bai and P. Perron, “Estimating and testing linear models with multiple structural changes,” *Econometrica*, vol. 66, pp. 47–78, 1998.

[22] J. Bai, “Likelihood ratio tests for multiple structural changes,” *Journal of Econometrics*, vol. 91, pp. 299–323, 1999.

[23] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin, “A statistical approach for array CGH data analysis,” *Bioinformatics*, vol. 6, no. 27, pp. 1–14, 2005.

[24] S. Li and R. Lund, “Multiple changepoint detection via genetic algorithms,” *Journal of Climate*, vol. 25, no. 2, pp. 674–686, 2012.

[25] J. Bai and P. Perron, “Computation and analysis of multiple structural change models,” *Journal of Applied Econometrics*, pp. 1–22, Feb. 2003.

[26] M. Lavielle, “Using penalized contrasts for the change-point problem,” *Institut National de Recherche en Informatique et en Automatique N° 5339*, Oct. 2004.

[27] Y. Guédon, “Explorating the segmentation space for the assessment of multiple change-points models,” *Institut National de Recherche en Informatique et en Automatique N° 6619*, Oct. 2008.

[28] G. Rigaill, “Pruned dynamic programming for optimal multiple change-point detection,” *arXiv:1004.0887v1*, Apr. 2010.

[29] J. Friedman, T. Hastie, and R. Tibshirani, “A note on the group Lasso and a sparse group Lasso,” *arXiv:1001.0736v1*, Jan. 2010.

[30] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, “A sparse-group Lasso,” *Journal of Computational and Graphical Statistics*, vol. 22, pp. 231–245, May 2013.

[31] R. Foygel and M. Drton, “Exact block-wise optimization in group Lasso and sparse group Lasso for linear regression,” *arXiv:1010.3320v2*, Nov. 2010.

[32] J. Fan, J. Lv, and L. Qi, “Sparse high dimensional models in economics,” *Annu Rev Econom.*, vol. 3, pp. 291 – 317, Sep. 2011.

[33] “Managing the deluge of ‘big data’ from space.” `http://www.jpl.nasa.gov/news/news.php?release=2013-299`, 2013.

[34] “Industrial network anomaly detection.” `http://global.ofweek.com/news/Industrial-network-anomaly-detection-23614`, 2015.

[35] P. Richtárik and M. Takáč, “Distributed coordinate descent method for learning with big data,” *Journal of Machine Learning Research*, Feb. 2016.

[36] Y. Zhang and L. Xiao, “Communication-efficient distributed optimization of self-concordant empirical loss,” *arXiv:1510.00263v1*, Jan. 2015.

[37] C. Ma and M. Takáč, “Partitioning data on features or samples in communication-efficient distributed optimization?,” *arXiv:1510.06688v1*, Oct. 2015.

[38] P. Richtárik and M. Takáč, “Parallel coordinate descent methods for big data optimization,” *Mathematical Programming*, vol. 156, p. 433484, Dec. 2016.

[39] Z. Harchaoui and C. *Ĺ*evy Leduc, “Multiple change-point estimation with a total variation penalty,” *Journal of the American Statistical Association*, vol. 105, pp. 1480–1493, Dec. 2010.

[40] J. Qian and L. Su, “Shrinkage estimation of regression models with multiple structural changes.” `http://jhqian.org/structure_change20130629.pdf`, Jun. 2013.

[41] D. Angelosante and G. Giannakis, “Group Lassoing change-points in piecewise-constant AR processes,” *EURASIP Journal on Advances in Signal Processing*, vol. 2012, p. 70, Mar. 2012.

[42] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, , and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices," *arXiv:1212.3753v3*, Jul. 2014.

[43] R. H. Loschi, J. G. Pontel, and F. R. B. Cruz, "Multiple change-point analysis for linear regression models," *Chilean Journal of Statistics*, vol. 1, pp. 93–112, Sep. 2010.

[44] D. Barry and J. A. Hartigan, "A Bayesian analysis for change point problems," *Journal of the American Statistical Association*, vol. 88, pp. 309–319, Mar. 1993.

[45] J. Hartigan, "Partition models," *Communications in Statistics - Theory and Methods*, vol. 19, pp. 2745–2756, Jun. 1989.

[46] M. Levorato and U. Mitra, "Fast anomaly detection in SmartGrids via sparse approximation theory," in *Proceedings of IEEE 7th Sensor and Multichannel Signal Processing Workshop*, (Hokoben, NJ), pp. 5–8, Jun. 2012.

[47] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated Gaussian designs," *Journal of Machine Learning Research*, vol. 11, pp. 2241–2259, Aug. 2010.

[48] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," *IEEE Trans. Inform. Theory*, vol. 57, pp. 6976–6994, Oct. 2011.

[49] Y. S. Soh and V. Chandrasekaran, "High-dimensional change-point estimation: Combining filtering with convex optimization," in *Proc. IEEE Intl. Symposium on Inform. Theory*, (Hong Kong, China), pp. 151 – 155, Jun. 2015.

[50] Y. Chi and Y. Wu, "Change-point estimation of high-dimensional streaming data via sketching," in *Proc. Asilomar Conf. on Signals, Systems and Computers*, (Pacific Grove, CA), pp. 102–106, Nov. 2015.

[51] G. K. Atia, "Change detection with compressive measurements," *IEEE Signal Processing Letters*, vol. 22, pp. 182–186, Feb. 2015.

[52] M. Jaggi, V. Smith, M. Takáč, J. Terhorst, S. Krishnan, T. Hofmann, and M. I. Jordan, "Communication-efficient distributed dual coordinate ascent," in *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems*, (Montreal, Canada), pp. 3068–3076, Dec. 2014.

[53] C. Ma, V. Smith, , M. Jaggi, M. I. Jordan, P. Richtárik, and M. Takáč, "Adding vs. averaging in distributed primal-dual optimization," in *Proceedings of the 32th International Conference on Machine Learning*, vol. 37, (Lille, France), pp. 3068–3076, Dec. 2015.

[54] O. Shamir, N. Srebro, and T. Zhang, "Communication-efficient distributed optimization using an approximate newton-type method," in *Proceedings of the 31 st International Conference on Machine Learning*, (Beijing, China), July 2014.

[55] F. Niu, B. Recht, C. Ré, and S. J. Wright, "Hogwild!: A lock-free approach to parallelizing stochastic gradient descent," in *NIPS'11 Proceedings of the 24th International Conference on Neural Information Processing Systems*, (Granada, Spain), Dec. 2011.

[56] J. Liu, S. J. Wright, C. Ré, V. Bittorf, and S. Sridhar, "An asynchronous parallel stochastic coordinate descent algorithm," in *Proceedings of the 31st International Conference on Machine Learning*, vol. 32, (Beijing, China), June 2014.

[57] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems*, (Stateline, NV), Dec. 2013.

[58] J. C. Duchi, M. I. Jordan, and H. B. McMaha, "Estimation, optimization, and parallelism when data is sparse," in *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems*, (Stateline, NV), Dec. 2013.

[59] M. Schmidt, N. L. Roux, and F. Bach, "A stochastic gradient method with an exponential convergence rate for finite training sets," in *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems*, (Stateline, NV), Dec. 2012.

[60] M. Schmidt, N. L. Roux, and F. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," *arXiv:1109.2415v2*, Dec. 2011.

[61] P. Machart, S. Anthoine, and L. Baldassarre, "Optimal computational trade-off of inexact proximal methods," *arXiv:1210.5034*, Oct. 2012.

[62] C. N. J. Ye Pu and M. N. Zeilinger, "Inexact alternating minimization algorithm for distributed optimization with an application to distributed mpc," *arXiv:1608.00413*, Aug. 2016.

[63] R. Tappenden, P. Richtárik, and J. Gondzio, "Inexact coordinate descent: Complexity and preconditioning," *Journal of Optimization Theory and Applications*, vol. 170, pp. 144–176, July 2016.

[64] B. Zhang, J. Geng, and L. Lai, "Multiple change-points estimation in linear regression models via sparse group Lasso," *IEEE Trans. Signal Processing*, vol. 63, pp. 2209 – 2224, May 2015.

[65] B. Zhang, J. Geng, and L. Lai, "Change-point estimation in high dimensional regression models," *IEEE Trans. Inform. Theory*, 2016. Submitted, available at users.wpi.edu/$\sim$bzhang.

[66] B. Zhang, J. Geng, W. Xu, and L. Lai, "Communication efficient distributed learning with feature partitioned data," *IEEE Trans. on Signal and Information Processing over Networks*, 2017. Submitted, available at users.wpi.edu/∼bzhang.

[67] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. The MIT Press, third ed., Jul. 2009.

[68] K. R. Davidson and A. P. Donsig, *Real Analysis and Applications-Theory in Practice*. New York: Springer, 2009.

[69] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[70] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, no. 1-4, pp. 259–268, 1992.

[71] D. M. Strong and T. F. Chan, "Exact solutions to total variation regularization problems," tech. rep., UCLA, 1996.

[72] S. Yang, J. Wang, W. Fan, X. Zhang, P. Wonka, and J. Ye, "An efficient admm algorithm for multidimensional anisotropic total variation regularization problems," in *SIGKDD*, (Chicago, USA), Aug. 2013.

[73] Z. Lu and L. Xiao, "On the complexity analysis of randomized block-coordinate descent methods," *arXiv:1305.4723v1*, May 2013.

[74] P. Richt$\acute{a}$rik and M. Tak$\acute{a}\check{c}$, "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function," *Mathematical Programming*, vol. 144, pp. 1–38, Apr. 2014.

[75] "NCEP/NCAR Reanalysis 1: Surface." `http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.surface.html`, 2014.

[76] R. F. Adler, G. J. Huffman, A. Chang, R. Ferraro, P.-P. Xie, J. Janowiak, B. Rudolf, U. Schneider, S. Curtis, D. Bolvin, A. Gruber, J. Susskind, P. Arkin, and E. Nelkin, "The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis," *Journal of Hydrometeorology Advanced Search*, vol. 4, pp. 1147–1167, Dec. 2003.

[77] B. Zhang, "R package ChangePointCalc." `http://users.wpi.edu/~bzhang/coding.html`. (R implementation of SGL for change-points estimation).

[78] Y. Nesterov, "Gradient methods for minimizing composite objective function," *Center for Operations Research and Econometrics Discussion Paper*, Sep. 2007.

[79] Y. Nesterov, *Introductory lectures on convex optimization: a basic course*. New York: Springer Science+Business Media, LLC, 2004.

[80] R. T. Rockafellar, *Convex Analysis*. Princeton, New Jersey: Princeton Univerisity Press, 1970.

[81] Y.-L. Yu, "Arithmetic duality for norms." `http://webdocs.cs.ualberta.ca/~yaoliang/mynotes/normduality.pdf`, 2012.