# Large Data Approaches to Thresholding Problems

by

Zhiyuan Lu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2019

Doctoral Committee:

Professor Moulinath Banerjee, co-Chair
Professor George Michailidis, co-Chair
Professor Matias D. Cattaneo
Professor Ya'acov Ritov

Zhiyuan Lu

jlnlu@umich.edu

ORCID iD: 0000-0003-1705-8885

# ACKNOWLEDGEMENTS

I thank my advisers and school for providing much learning experience during my years as a PhD student, and my parents for their support.

# TABLE OF CONTENTS

# LIST OF FIGURES

**Figure**

# LIST OF TABLES

**Table**

# ABSTRACT

Statistical models with discontinuities have seen much use in a variety of situations, in practical fields such as statistical process control, processing gene data, and econometrics. The study of such models is usually concerned with locating the these discontinuities, which methodologically cause various issues as estimation requires nonstandard optimization problems. With the contemporary increase in computer power and memory, it becomes more relevant to view these problems in the context of very large datasets, a context which introduces further complications for estimation. In this thesis, we study two major topics in threshold estimation, with models, methodology, and results motivated by the concern towards handling big data.

Our first topic focuses on the change point problem, which involves detection of the locations where a change in distribution occurs within a data sequence. A variety of methods have been proposed and studied in this area, with novel approaches in the case where the number of change points is an unknown that could be greater than 1, making exhaustive search methods infeasible.

Our contribution in this problem is motivated by the principle that only the data points close to the change points are useful for their estimation while other points are extraneous. From this observation we propose a zoom in estimation method which

efficiently subsamples the data for estimation while not compromising the accuracy. The resulting method runs in sublinear time, while existing methods all run in linear time or above. Furthermore, the nature of this new methodology allows us to characterize the asymptotic distribution even in the case where the number of change point parameters increases without bound, a type of result not replicated in this field.

The second topic regards the change plane model, which involves a real valued signal over a multiple dimensional space with a discontinuity delineated by a hyperplane. Practically the change plane model is used to combine regression between a covariate and response variable, while performing unsupervised classification onto the covariate. As change -plane models in growing dimensions have not been studied in the literature, we confine ourselves to canonical models in this dissertation, as a first approach to these problems. in terms of details, we establish fundamental convergence and support selection properties (the latter for the high-dimensional case) and present some simulation results.

# CHAPTER I

# Introduction

In this document we will focus on estimation of discontinuities in various models arising from two main categories, change point models which focuses on discontinuities at points in one-dimensional sequences of data, and change plane models focused on estimating the discontinuity at a hyper-plane in higher dimensional data. These two types of models are utilized in different settings in biology, econometrics, and other fields: change point detection is concerned with detecting the actual discontinuities in one dimensional data sequences while change point models are concerned with delineating the sub-populations in covariate-response models. Nonetheless, they are mathematically similar in nature, and in this document we study both, with a common concern in how to handle large data sets.

First, in Chapter II, we focus on the topic of change points in data sequences. As a field of research, this topic is usually concerned with a model where the data is sequentially ordered and has a distribution which is fixed along segments but sees sudden changes at a few points (the change points). Alternatively, one can see the change point problem as the estimation of discontinuities in a signal of data, making it a non-standard problem in terms of actual estimation. Consequently, even the fastest change point estimation methods involving searching along the entire sequence

and calculating an expression at each point, with some methods becoming very time consuming when considering multiple change points within the data.

Our contribution to this field is a method which alleviates this problem, especially for longer data sequences. The new method introduced here stem from the following intuition: only points close to the actual change points are useful for their estimation, while data points further away are essentially extraneous. From this idea we develop a multi-stage subsampling scheme which first roughly locate the change points to small intervals before estimating within these localized neighborhoods. The general result is that this method does consistently estimate for the change points with a lower order of computational time than using the full data set, while at the same time not sacrificing the accuracy of the estimate despite not utilizing every point.

To illustrate the basic premise of our method and its asymptotic properties, we initially work with the simplest change point model: the stump model with only a single change point, and only the mean of the distribution changes. In Section 2.2 we describe how to apply the intelligent sampling procedure to the single change point and derive its properties. Among our results, we show that the rate of convergence and even the asymptotic distribution are the same as the estimator using the full data set. Besides theoretical results, we also perform an analysis on the computational order, which scales to almost square root of the data size, compared to scaling linearly with the data size if one estimates using the full data. Proofs for this section can be found in Section A.1.1, where a more general model was studied.

We next turn our attention to a more sophisticated model that is more pertinent to application and current research, the multiple change model where both the number and location of the change points are unknown and need to be estimated. In

Section 2.3, we consider a canonical model analyzed often in change point literature. Next, in Section 2.3.1, we describe our intelligent sampling procedure when applied to this model, followed by asymptotic results. In terms of the latter, not only were we able to derive the rate of convergence for the estimators and their asymptotic distribution in a finite change point framework, which are common results found in literature, but in Theorem II.3 we characterize the distribution of the estimators even should the number of change points grow without bound. This unique result makes our procedure useful for statistical inference. Proofs for this section can be found in Section A.2, and supplementary material useful for deriving Theorem II.3 can be found in Section A.3.

Next, in Section 2.4, we focus on methodological and computational aspects of intelligent sampling on multiple change points, which are more complex than for the single change point problem. First, we describe two techniques in Section 2.4.1 and 2.4.2 which can be used to complete the procedures in Section 2.3.1 into a fully implementable procedure. An analysis on the computational cost, which now depends on the number of change points present and their spacing, follows in Section 2.4.3, followed by an illustration of computational savings in Section 2.5. As a demonstrate of implementing intelligent sampling, we close the chapter with various simulation results and an analysis on an internet traffic dataset in Section 2.7 and 2.8.

In the next chapter, we focus on the change plain problem, which is also concerned with the estimation of discontinuities, but in a higher dimensional framework rather than on data sequences. These models usually involve a covariate-response setup where the relationship between $X$ and $Y$ features a discontinuity delineated by a hyperplane. In the spirit of contemporary work, for us to focus on large data sets, it would be appropriate to consider the case where the dimension of the problem

grows with the sample size. As existing work on the change point problem focuses on the fixed dimensional framework, we focus on establishing results for basic canonical models.

To illustrate the basic asymptotic properties for growing dimensions, in Section 3.2 we first analyze this setting with the least squares estimator on a basic model with a dimension that is growing but much less than the sample size. In that section we establish the convergence rate and demonstrate that it is essentially min-max optimal. Unfortunately, the least squares estimator has certain undesirable properties (such as require the estimation of nuisance parameters) so we next introduce an alternative estimator without such flaws in 3.4.

The analysis of this estimator will continue on the high dimensional (dimension greater than sample size) in Section 3.5. In addition to deriving consistency and rate of convergence, we were also able to derive consistency in terms of recovering the support set of our estimator.

# CHAPTER II

# Change Points in Data Sequences

## 2.1 Introduction

Change point analysis has been extensively studied in both the statistics and econometrics literature *Basseville and Nikiforov* (1993), due to its wide applicability in many application areas, including economics and finance *Frisén* (2008), quality control *Qiu* (2013), neuroscience *Koepcke et al.* (2016), etc. A change point represents a discontinuity in the parameters of the data generating process. The literature has investigated both the *offline* and *online* versions of the problem *Basseville and Nikiforov* (1993); *Csörgö and Horváth* (1997). In the former case, one is given a sequence of observations and questions of interest include: (i) whether there exists a change point and (ii) if there exists one (or multiple) change point(s), identify its (their) location, as well as estimate the parameters of the data generating process to the left and right of it (them). In the latter case, one is obtaining new observations in a sequential manner and the main interest is in quickest detection of the change point.

In the era of big data, researchers are faced with large numbers of observations whose processing and storage pose a number of computational challenges. For example, monitoring traffic in high-speed computer and communication networks and trying to identify shifting user patterns, as well as malicious activities is of great inter-

est to network engineers (see *Kallitsis et al.* (2016) and references therein). However, the number of observations produced corresponding to number of packets or their corresponding payload in bytes at a granular temporal scale is in the thousands per minute. Nevertheless, it is of value to network operators to examine these traces and if something of interest is identified, to store the appropriate segment for further analysis. Analogous problems come up in other systems including manufacturing processes *Shen et al.* (2016), or other cyber-physical systems equipped with sensors *Khan et al.* (2016). This exercise requires temporary storage of the data, possibly in a distributed manner, as well as doing the necessary calculations on them.

In this chapter, we address the offline problem for *extremely long* data sequences, the analysis of which by conventional modes of analysis turns out to be prohibitive owing to the massive size involved. The identification of change-points in a sequence, which constitute local discontinuities, requires some sort of a search procedure. A list of such methods, along with summaries of their results, can be found in *Niu et al.* (2016). Exhaustive search is typically infeasible with multiple change-points, since the search-time grows exponentially in the number of change-points. A variety of intelligent methods have been proposed in the literature, e.g. binary segmentation (see *Venkatraman* (1992) and *Fryzlewicz et al.* (2014)), wild binary segmentation (*Fryzlewicz et al.* (2014)), multi-scale methods (*Frick et al.* (2014) and *Pein et al.* (2016)), $\ell_1$ penalization (*Huang et al.* (2005) and *Harchaoui and Lévy-Leduc* (2010)); however, the best feasible computation time is of the order $N$ or $N \log N$, which, while being reasonable, is still a huge ask *when one is considering a sequence of observations of length in the tens or hundreds of millions or even larger.*

The point of view we adopt is the following: we are given a data sequence of massive length and are interested in identifying the major changes in this series. While *short-lived* changes of various intensities in mean shifts may occur frequently, such transient perturbations do not affect the overall performance of most engineered

or physical systems. Hence, it is reasonable to assume that the number of truly significant changes that *persist* over time, is not particularly large. However, sifting through the entire mass of data to detect those changes is computationally expensive and on many occasions even prohibitive. The modeling strategy uses a very long piece-wise constant mean model with multiple jumps – where the number of jumps increases at a rate much slower than the length of the series — that are not too small relative to the fluctuations induced by noise. The 'piece-wise constant with jumps' strategy has been considered by a variety of authors (see *Niu et al.* (2016) and references therein) for the problem at hand and is convenient for developing the methodology. Further, as the emulation experiment in Section 2.8 demonstrates, our proposed method consistently picks up the persistent structural changes in the presence of multiple spiky signals while staying agnostic to the latter. Nevertheless, if short duration shifts are also of interest, the interested researcher can further analyze the persistent segments (which are of smaller order than $N$) with any of the available procedures in the literature in a *parallel* fashion, thus retaining the computational gains achieved by intelligent sampling.

We now articulate the contributions of this chapter.

**1.** Our *key objective* is to propose an effective solution to the computational problem discussed above via a strategy called "intelligent sampling", which proceeds by making two (or more) passes over the time-series at different levels of resolution. The first pass is carried out at a coarse time resolution and enables the analyst to obtain pilot estimates of the change-points, while the second investigates stretches of the time-series in (relatively) small neighborhoods of the initial estimates to produce updated estimates enjoying a high level of precision; in fact, essentially the same level of precision as would be achieved by an analysis of the entire massive time-series. The core advantage of our proposed method is that it reduces computational time from *linear* to *sub-linear* under appropriate conditions, and, in fact close to square-root

order, if the number of change-points is small relative to the length of the time-series. It is established that the computational gains (and analogously other processing gains from input-output operations) can be achieved *without compromising* any statistical efficiency.

**2.** Most of our results will be rigorously developed in the signal plus noise model with independent and identically distributed (iid) normal errors, which has also been used by other authors in this area to showcase their research, e.g. *Fryzlewicz et al.* (2014), *Niu and Zhang* (2012), and *Zhang and Siegmund* (2007), as it an attractive canonical model and amenable to theoretical analysis. However, we also provide results/indicate extensions when the errors exhibit short or long-range dependence and in the presence of non-stationarity, since both these features are likely to be present with very long data sequences. Some empirical evidence is provided to this effect. We do not study non-Gaussian errors in full generality but develop some extensions in such directions, see, Remark 5 and the section in the supplement referred therein, and also Section 2.9 for some comments on this issue.

Further, the focus of the presentation is on 2-stage procedures that provide all the key insights into the workings of the intelligent sampling procedure. However, for settings where the size of the data exceeds $10^{10}$, multiple stages are required to bring down the analyzed subsample to a manageable size. We therefore cover extensions to multi-stage intelligent sampling procedures as well as address how samples should be allocated at these different stages. Furthermore, such massive data sequences, often, cannot be effectively stored in a single location. This does not pose a problem for intelligent sampling as it adapts well to distributed computing: it can be applied on the reduced size subsamples at the various locations where the original data are stored, followed by a single subsequent back and forth communication between the various locations and the central server, and subsequent calculations essentially carried out on the local servers. This is elaborated on in Section 2.5.

**3.** On the inferential front, we establish asymptotic approximations to the joint distribution of the change-point estimates obtained by intelligent sampling in terms of the distribution functions of symmetric drifted random walks on the set of integers [Theorems II.3 and II.5], which can then be used to provide explicitly computable asymptotic joint confidence intervals for both finitely many and a *growing number* of change points. While concentration properties of change-point estimates around the the true parameters in multiple change point problems are known, to the best of our knowledge, such results involve hard-to-pin-down constants (this is discussed in some more detail in Section 2.4.2 in the context of binary segmentation) and are therefore difficult to use in practical settings. Our prescribed methods involve estimating signal to noise ratios at the different change-points, which is easy to accomplish, and values of quantiles of symmetric Gaussian random walks for different values of the drift parameter (which can be pre-generated on a computer). The arguments involved in establishing Theorem II.3 require, among other things, careful analyses of the distribution functions of both symmetric and asymmetric Gaussian random walks (see part B of Supplement, Section A.2) which may very well prove useful in many other contexts.

The remainder of this chapter is organized as follows. Section 2.2 addresses intelligent sampling for the simpler *single* change point problem, which provides some fundamental insights into the nature of the procedure and its theoretical and computational properties. Section 2.3 deals with the main topic of this study: intelligent sampling for the multiple change point problem with a growing number of change-points, and presents the main theoretical results of the section. Section 2.4 develops the practical methodology for intelligent sampling using *binary segmentation* as the working procedure at Stage 1, and studies the computational complexity of the resulting approach. Section 2.5 provides an elaborate study of the minimum subsample size required for precise inferences as a function of the length of the full data sequence

and the signal-to-noise ratio using multiple stage procedures. Extensions to non-iid settings which are more pertinent for the more special case of time-series data are discussed in Section 2.6. The numerical performance of the procedure is illustrated on synthetic data in Section 2.7, while an application to real Internet data is presented in 2.8. Section 2.9 concludes with a discussion of possible extensions of the intelligent sampling procedure, both in terms of alternative Stage 1 procedures (like wild binary segmentation and SMUCE), and also to other kinds of data (non-Gaussian data, discrete data, decaying signals). In the appendix, the complete proofs of of all results, the proofs of accompanying technical lemmas, and elaborate discussions of various facets of intelligent sampling are presented.

## 2.2 Intelligent Sampling for the Single Change Point Problem

### 2.2.1 Single Change Point Model

The simplest possible setting for the change point problem is the *stump* model, where data $(1/N, Y_1), \cdots, (N/N, Y_N)$ are available with $Y_i = f(X_i) + \varepsilon_i$ for $i = 1, \cdots, N$, and where the error term $\varepsilon_i$ is independent and identically distributed (iid) following a $N(0, \sigma^2)$ distribution, while the function $f$ takes the form

$$f(x) = \alpha \cdot 1(x \leq \tau) + \beta \cdot 1(x > \tau), \qquad x \in (0, 1), \tag{2.1}$$

for some constants $\alpha, \beta \in \mathbb{R}$, $\alpha \neq \beta$, and $\tau \in (0, 1)$: the so-called 'stump' model. For estimating the *change point* $\tau$ we employ a least squares criterion, given by

$$(\hat{\alpha}, \hat{\beta}, \hat{\tau}) := \underset{(a,b,t)\in\mathbb{R}^2\times(0,1)}{\arg\min} \sum_{i=1}^{N} \left(Y_i - a \cdot 1(i/N \leq t) - b \cdot 1(i/N > t)\right)^2. \tag{2.2}$$

Using techniques similar to those in Section 14.5 of *Kosorok* (2007), we can estab-

lish that the estimator $\hat{\tau}$ is consistent for $\tau_N := \lfloor N\tau \rfloor / N$, which acts as the change point among the covariates lying on the even grid.

**Proposition 1.** *For the stump model with normal errors the following hold:*
*(i) Both* $(\hat{\alpha} - \alpha)$ *and* $(\hat{\beta} - \beta)$ *converge to 0 with rate* $O_p(N^{-1/2})$.
*(ii) The change point estimate* $\hat{\tau}$ *satisfies*

$$\mathbb{P}\left[N(\hat{\tau} - \tau_N) = k\right] \to \mathbb{P}[L = k] \text{ for all } k \in \mathbb{Z} \tag{2.3}$$

*where* $L = \underset{i \in \mathbb{Z}}{\arg\min}\, X(i)$, *and the random walk* $\{X(i)\}_{i \in \mathbb{Z}}$ *is defined as*

$$X(i) = \begin{cases} \Delta(\varepsilon_1^* + ... + \varepsilon_{i-1}^* + \varepsilon_i^*) + i\Delta^2/2, & i > 0 \\ 0, & i = 0 \\ -\Delta(\varepsilon_{i+1}^* + ... + \varepsilon_{-1}^* + \varepsilon_0^*) + |i|\Delta^2/2, & i < 0, \end{cases} \tag{2.4}$$

*with* $\varepsilon_0^*, \varepsilon_1^*, \varepsilon_2^*, \ldots$ *and* $\varepsilon_{-1}^*, \varepsilon_{-2}^*, \ldots$ *being iid* $N(0, \sigma^2)$ *random variables and* $\Delta := \beta - \alpha$.

Next, we make several notes on the random variable $L$ introduced in the above proposition, as it appears multiple times throughout the remainder of this paper. Although, at a glance, the distribution of $L$ depends on two parameters, $\Delta$ and $\sigma$, in actuality $L$ is completely determined by the signal-to-noise ratio $\Delta/\sigma$ due to the Gaussian setting. To see this, note that we can re-write $L = \underset{i \in \mathbb{Z}}{\arg\min}(Z_i/|\Delta\sigma|)$ where

$$\frac{X(i)}{|\Delta\sigma|} = \begin{cases} \text{sgn}(\Delta)(\varepsilon_1^*/\sigma + \cdots + \varepsilon_i^*/\sigma) + i|\Delta/\sigma|/2, & i > 0 \\ 0, & i = 0 \\ -\text{sgn}(\Delta)(\varepsilon_{i+1}^*/\sigma + \cdots + \varepsilon_0^*/\sigma) + |i\Delta/\sigma|/2, & i < 0 \end{cases} \tag{2.5}$$

Since $\{\text{sgn}(\Delta)\varepsilon_i^*/\sigma\}_{i \in \mathbb{Z}}$ are iid $N(0,1)$ random variables, invariant under $\Delta$ and $\sigma$, it follows that $L$ only depends on the single parameter $\Delta/\sigma$. Hence, from here on,

denote the associated random process as

$$X_\Delta(i) = \begin{cases} \text{sgn}(\Delta)(\varepsilon_1^* + \cdots + \varepsilon_i^*) + i|\Delta|/2, & i > 0 \\ 0, & i = 0 \\ -\text{sgn}(\Delta)(\varepsilon_{i+1}^* + \cdots + \varepsilon_0^*) + |i| \cdot |\Delta|/2, & i < 0 \end{cases} \tag{2.6}$$

where $\varepsilon_j^*$ for $j \in \mathbb{Z}$ are all iid $N(0,1)$ random variables. Denote the argmin of the random walk $X_\Delta(i)$ as $L_\Delta = \arg\min_{i \in \mathbb{Z}} X_\Delta(i)$. An immediate observable property of $L_\Delta$ is the stochastic ordering with respect to $|\Delta|$:

**Proposition 2.** *Suppose we have constants $\Delta_1, \Delta_2 \in \mathbb{R}$ such that $0 < |\Delta_1| < |\Delta_2|$, then for any positive integer $k$*

$$\mathbb{P}[|L_{\Delta_1}| \leq k] \leq \mathbb{P}[|L_{\Delta_2}| \leq k] \tag{2.7}$$

Practically, this stochastic ordering implies that if the $1-\alpha$ quantile $Q_{\Delta_1}(1-\alpha)$ of $|L_{\Delta_1}|$ is known, then $Q_{\Delta_1}(1-\alpha)$ can also serve as a conservative $1-\alpha$ quantile of $|L_{\Delta_2}|$ for any $|\Delta_2| \geq |\Delta_1|$. This can be useful in settings where given $J > 0$ random variables $L_{\Delta_i}$ for $i = 1 \ldots, J$, we desire positive integers $\ell_i$ for $i = 1, \ldots, J$ such that $\mathbb{P}[|L_{\Delta_i}| \leq \ell_i] \geq 1-\alpha$ for $i = 1, \ldots, J$. This scenario will appear in later sections where we consider models containing several change points with possibly different jump sizes. In such situations, a simple solution is to take $\ell_i = Q_{\Delta_m}(1-\alpha)$ for all $i$ where $m = \arg\min_{1,\ldots,J} |\Delta_i|$, or in other words letting each $\ell_i$ be the $1-\alpha$ quantile of the $|L_{\Delta_i}|$ with the smallest parameter. Alternatively we can generate a table of quantiles for distributions $L_{\delta_1}, L_{\delta_2}, L_{\delta_3}, \ldots$ for a mesh of positive constants $\delta_1 < \delta_2 < \ldots$ (e.g. we can let the $\delta_j = 0.1j$ for $j = 5, \ldots, 1000$), and let $\ell_i = Q_{\delta_j}(1-\alpha)$ where $\delta_j = \max\{\delta_k : \delta_k \leq \Delta_i\}$, for $i = 1, \ldots, J$.

### 2.2.2 The Intelligent Sampling Procedure and its Properties

(ISS1): From the full data set of $\left(\frac{1}{N}, Y_1\right), \left(\frac{2}{N}, Y_2\right), ..., (1, Y_N)$, take an evenly spaced subsample of approximately size $N_1 = K_1 N^{\gamma}$ for some $\gamma \in (0,1)$, $K_1 > 0$: thus, the data points are $\left(\frac{\lfloor N/N_1 \rfloor}{N}, Y_{\lfloor N/N_1 \rfloor}\right)$, $\left(\frac{2\lfloor N/N_1 \rfloor}{N}, Y_{2\lfloor N/N_1 \rfloor}\right)$, $\left(\frac{3\lfloor N/N_1 \rfloor}{N}, Y_{3\lfloor N/N_1 \rfloor}\right)$ ...

(ISS2): On this subsample apply least squares to obtain estimates $\left(\hat{\alpha}^{(1)}, \hat{\beta}^{(1)}, \hat{\tau}_N^{(1)}\right)$ for parameters $(\alpha, \beta, \tau_N)$.

By the results for the single change-point problem presented above, $\hat{\tau}_N^{(1)} - \tau_N$ is $O_p(N^{-\gamma})$. Therefore, if we take $w(N) = K_2 N^{-\gamma+\delta}$ for some small $\delta > 0$ (much smaller than $\gamma$) and any constant $K_2 > 0$, with probability increasing to 1, $\tau_N \in [\hat{\tau}_N^{(1)} - w(N), \hat{\tau}_N^{(1)} + w(N)]$. In other words, this provides a neighborhood around the true change point as desired; hence, in the next stage only points within this interval will be used.

(ISS3): Fix a small constant $\delta > 0$. Consider all $i/N$ such that $i/N \in [\hat{\tau}_N^{(1)} - K_2 N^{-\gamma+\delta}, \hat{\tau}_N^{(1)} + K_2 N^{-\gamma+\delta}]$ and $(i/N, Y_i)$ was not used in the first subsample. Denote the set of all such points as $S^{(2)}$.

(ISS4): Fit a step function on this second subsample by minimizing

$$\sum_{i/N \in S^{(2)}} \left(Y_i - \hat{\alpha}^{(1)} 1(i/N \leq d) - \hat{\beta}^{(1)} 1(i/N > d)\right)^2$$

with respect to $d$, and take the minimizing $d$ to be the second stage change point estimate $\hat{\tau}_N^{(2)}$.

The next theorem establishes that the intelligent sampling estimator $\hat{\tau}_N^{(2)}$ is *consistent with the same rate of convergence* as the *estimator based on the full data*.

**Theorem II.1.** *For the stump single change point model, the estimator obtained based on intelligent sampling satisfies*

$$|\hat{\tau}_N^{(2)} - \tau_N| = O_p(1/N).$$

*Proof.* See Section A.1.3 in Supplement Part A, where a result for a more general model is proven. □

To make a clean statement of the asymptotic distribution, we introduce a slight modification to the definition of the true change point and define a new type of 'distance' function $\lambda_2 : [0,1]^2 \mapsto \mathbb{Z}$, as follows. First, for convenience, denote the set of $X_i$'s of the first stage subsample as

$$S^{(1)} := \left\{ \frac{i}{N} : i \in \mathbb{N},\, i < N,\, i \text{ is divisible by } \lfloor N/N_1 \rfloor \right\}, \tag{2.8}$$

then for any $a, b \in (0,1)$

$$\lambda_2(a,b) := \begin{cases} \sum_{i=1}^{N} 1\left(a < \frac{i}{N} \le b,\, \frac{i}{N} \notin S^{(1)}\right) & \text{if } a \le b\,, \\ -\sum_{i=1}^{N} 1\left(b < \frac{i}{N} \le a,\, \frac{i}{N} \notin S^{(1)}\right) & \text{otherwise.} \end{cases} \tag{2.9}$$

The modified 'distance' $\hat{\tau}_N$ is $\lambda_2(\tau_N, \hat{\tau}_N^{(2)})$, instead of $N(\hat{\tau}_N^{(2)} - \tau_N)$, does converge weakly to a distribution.

**Theorem II.2.** *For any integer $\ell$,*

$$\mathbb{P}\left[\lambda_2\left(\tau_N, \hat{\tau}_N^{(2)}\right) = \ell\right] \;\to\; \mathbb{P}[L_{\Delta/\sigma} = \ell]\,. \tag{2.10}$$

*Proof.* See Section A.1.4 in Supplement Part A where this is shown for a more general model. □

**Computational gains:** The results above establish that the two stage procedure can, using a subset of the full data, be asymptotically almost as precise as employing the full dataset. In practice this allows for quicker estimation of big datasets without losing precision. The first stage uses about $N_1 \sim N^\gamma$ points to perform least squares fitting of a stump model, and this step takes $O(N^\gamma)$ computational time. The second stage applies a least-squares fit of a step function on the set $S^{(2)}$, which contains $O(N^{1-\gamma+\delta})$ points and therefore uses $O(N^{1-\gamma+\delta})$ time.

Hence, the two stage procedure requires order $N^\gamma \vee N^{1-\gamma+\delta}$ computation time, which is minimized by setting $\gamma = 1-\gamma+\delta$, or $\gamma = \frac{1+\delta}{2}$. As $\delta$ tends to 0 (any small positive value of $\delta$ yields the above asymptotic results), the optimal $\gamma$ tends to 1/2. Therefore, one should employ $N_1 = \sqrt{N}$ at the first stage and the second stage sample should be all points in the interval $[\hat{\tau}_N^{(1)} - K_2\sqrt{N}, \hat{\tau}_N^{(1)} + K_2\sqrt{N}]$, minus those at the first stage, where $K_2$ ensures that this interval contains $\tau$ with an acceptable high probability $1-\alpha$. If one knows the jump size $\Delta$, $K_2$ can be determined as the $1-\frac{\alpha}{2}$ quantile of the random variable $L_{\Delta/\sigma}$; in the realistic unknown $\Delta$ case, a lower estimate of $\Delta$ can yield a corresponding conservative value of $K_2$.

**Remark 1.** *The $\lambda_2$ distance was introduced above because it is generally not possible to derive an asymptotic distribution for $N(\hat{\tau}_N^{(2)} - \tau_N)$. Indeed, one can manufacture parameter settings quite easily, that produce different limit distributions along different subsequences. For a specific example, see Remark 19 in Supplement Part A.*

**Remark 2.** *We can extend our 2-stage procedure by adding in more stages. In the 2-stage version, we first use a subsample of size $N^\gamma$ to find some interval $[\hat{\tau}^{(1)} - K_1 N^{-\gamma+\delta_1}, \hat{\tau}^{(1)} + K_1 N^{-\gamma+\delta_1}]$ which contains the true value of $\tau$ with probability going to 1. However, nothing forces us to use all the data contained within this interval at the second stage. We can, instead, apply a two stage procedure on this interval as well, that is, take a subset of $N^\zeta$ (for some $0 < \zeta < 1-\gamma+\delta_1$) points from the second stage interval for estimation. From this, we obtain a second stage estimate $\hat{\tau}^{(2)}$ and*

*an interval* $[\hat{\tau}^{(2)} - K_2 N^{-\gamma-\zeta+\delta_1+\delta_2}, \hat{\tau}^{(2)} + K_2 N^{-\gamma-\zeta+\delta_1+\delta_2}]$ *(note that* $\delta_1$ *and* $\delta_2$ *can be as small as one pleases) which contains* $\tau$ *with probability going to 1. In the third stage, we finally take all points in the aforementioned interval (leaving aside those used in previous stages) to obtain an estimate* $\hat{\tau}^{(3)}$*.*

*Such a procedure will have the same rate of convergence as the one using the full data:* $(\hat{\tau}^{(3)} - \tau) = O_p(1/N)$*, and the same asymptotic distribution (in terms of a "third stage distance" similar to how* $\lambda_2$ *was defined) as the one and two stage procedures. In terms of computational time, the first stage takes* $O(N^\gamma)$ *time, the second stage* $O(N^\zeta)$ *time, and the final stage* $O(N^{1-\gamma-\zeta+\delta_1+\delta_2})$ *time, for a total of* $O((N^\gamma \vee N^\zeta \vee N^{1-\gamma-\zeta+\delta_1+\delta_2}))$ *time, which can reach almost* $O(N^{1/3})$ *time. In general, a* $k$ *stage procedure, which works along the same lines can operate in almost as low as* $O(N^{1/k})$ *time.*

## 2.3 The Case of Multiple Change Points

Suppose one has access to a data set $Y_1, Y_2, \ldots, Y_N$ generated according to the following model:

$$Y_i = \theta_i + \varepsilon_i, \qquad i = 1, 2, 3, ..., N \tag{2.11}$$

where the $\theta_i$'s form a piecewise constant sequence for any fixed $N$ and the $\varepsilon_i$'s are zero-mean error terms[1]. The signal is flat apart from jumps at some unknown change points $1 = \tau_0 < \tau_1 < ... < \tau_J < \tau_{J+1} = N$: i.e. $\theta_{i_1} = \theta_{i_2}$ whenever $i_1, i_2 \in (\tau_j, \tau_{j+1}]$ for some $j \in \{0, ..., J\}$. The number of change points $J = J(N)$ is also unknown and needs to be estimated from the data. We impose the following basic restrictions on this model:

[1]To be more precise we consider the triangular array of sequences $\theta_{i,N}$, which are piecewise constant in $i$. The error terms $\varepsilon_i = \varepsilon_{i,N}$ also form a triangular array, but we suppress the notation for brevity

(M1): there exists a constant $\bar{\theta} \in (0, \infty)$ not dependent on $N$, such that $\max_{i=1,...,N} |\theta_i| \leq \bar{\theta}$;

(M2): there exists a constant $\underline{\Delta}$ not dependent on $N$, such that $\min_{i=0,...,J} \left|\theta_{\tau_{i+1}} - \theta_{\tau_i}\right| \geq \underline{\Delta}$;

(M3): there exists a $\Xi \in [0, 1)$ and some $C > 0$, such that $\delta_N := \min_{i=0,...,J} (\tau_{i+1} - \tau_i) \geq C N^{1-\Xi}$ for all large $N$;

(M4): $\varepsilon_i$ for $i = 1, ..., N$ are i.i.d. $N(0, \sigma^2)$.

**Remark 3.** *The second assumption above stipulates that the minimum gap between two consecutive stretches is bounded away from 0. In the context of identifying long and significantly well-separated persistent stretches in a big data setting, this is a reasonable assumption.*

### 2.3.1 Intelligent Sampling on Multiple Change Points

The intelligent sampling procedure in the multiple change-points case works in two (or more) stages: in the two-stage version, as in Section 2.2, the first stage aims to find rough estimates of the change points using a uniform subsample (Steps ISM1-ISM4) and the second stage produces the final estimates (Steps ISM5 and ISM6).

(ISM1): Start with a data set $Y_1, \cdots, Y_N$ described in (2.11).

(ISM2): Take $N_1 = K_1 N^\gamma$ for some $K_1$ and $\gamma \in (\Xi, 1)$ such that $N/N_1 = o(\delta_N)$; for $j = 1, \cdots, N^*$ where $N^* := \left\lfloor \frac{N}{\lfloor N/N_1 \rfloor} \right\rfloor$, consider the subsample $\{Z_j\} = \{Y_{j\lfloor N/N_1 \rfloor}\}$.

The subsample $Z_1, Z_2, ...$ can also be considered a data sequence structured as in (2.11), and since $\delta_N >> N/N_1$, there are jumps in the signal at $\tau_j^* := \left\lfloor \frac{\tau_j}{\lfloor N/N_1 \rfloor} \right\rfloor$ for

$j = 1, ..., J$, with corresponding minimum spacing

$$\delta^*_{N^*} := \min_{i=1,...,J+1} |\tau^*_i - \tau^*_{i-1}|$$
$$= \frac{1}{\lfloor N/N_1 \rfloor} \left( \min_{i=1,\cdots,J+1} |\tau_i - \tau_{i-1}| + O(1) \right) = \left( \frac{N_1}{N} \delta_N \right) (1 + o(1)) . \quad (2.12)$$

(ISM3): Apply some multiple change point estimation procedure (such as binary segmentation) to the set of $Z_i$'s to obtain estimates $\hat{\tau}^*_1, ..., \hat{\tau}^*_{\hat{J}}$ for the $\tau^*_i$s and $\hat{\nu}^{(1)}_0, \ldots, \hat{\nu}^{(1)}_{\hat{J}}$ for the levels $(\nu_0, \nu_1, \ldots, \nu_J) = (\theta_1, \theta_{\tau_1+1}, \theta_{\tau_2+1}, \ldots, \theta_{\tau_J+1})$.

- the choice of the procedure does not matter so long as the estimates satisfy

$$\mathbb{P}\left[ \hat{J} = J, \max_{i=1,...,J} |\hat{\tau}^*_i - \tau^*_i| \leq w^*(N^*), \max_{i=0,...,J} |\hat{\nu}^{(1)}_i - \nu_i| \leq \rho_N \right] \to 1 \quad (2.13)$$

for some sequence $w^*(N^*)$ such that $w^*(N^*) \to \infty$, $w^*(N^*) = o(\delta^*_{N^*})$ and $\rho_N \to 0$.

(ISM4): Convert these into estimates for the $\tau_i$'s by letting $\hat{\tau}^{(1)}_j := \hat{\tau}^*_j \lfloor N/N_1 \rfloor$ for $j = 1, ..., \hat{J}$.

- taking $w(N) := (w^*(N^*) + 1)\lfloor N/N_1 \rfloor$, expression (2.13) gives

$$\mathbb{P}\left[ \hat{J} = J, \max_{i=1,...,J} |\hat{\tau}^{(1)}_i - \tau_i| \leq w(N), \max_{i=0,...,J} |\hat{\nu}^{(1)}_i - \nu_i| \leq \rho_N \right] \to 1. \quad (2.14)$$

- as a consequence of conditions in (ISM3), $w(N) \to \infty$, $w(N) = o(\delta_N)$, and $w(N) > CN^{1-\gamma}$ for some constant $C$.

(ISM5): Fix any integer $K > 1$, and consider the intervals $\left[ \hat{\tau}^{(1)}_i - Kw(N), \hat{\tau}^{(1)}_i + Kw(N) \right]$ for $i = 1, ..., \hat{J}$. Denote by $S^{(2)}\left( \hat{\tau}^{(1)}_i \right)$ all integers in this interval not divisible by $\lfloor N/N_1 \rfloor$.

(ISM6): For each $i = 1, ..., \hat{J}$, let

$$\hat{\tau}_i^{(2)} = \underset{d \in S^{(2)}\left(\hat{\tau}_i^{(1)}\right)}{\arg\min} \left( \sum_{j \in S^{(2)}\left(\hat{\tau}_i^{(1)}\right)} \left[ Y_j - (\hat{\nu}_{i-1}^{(1)} 1(j < d) + \hat{\nu}_i^{(1)} 1(j \geq d)) \right]^2 \right) . \quad (2.15)$$

**Remark 4.** *As with the single change point problem, a $p > 2$ stage procedure can be constructed. This would involve steps (ISM1) to (ISM5), but afterwards a $p - 1$ stage procedure as described in Remark 2 for estimating single change points will be applied on every interval $[\hat{\tau}_i \pm Kw(N)]$. To illustrate our main points we will show results for the two stage procedure, but some numerical results are provided later for multi-stage procedures.*

The intervals $[\hat{\tau}_j^{(1)} \pm Kw(N)]$ referred to at stage ISM4 all respectively contain $[\tau_j \pm (K-1)w(N)]$ with probability going to 1. These latter intervals have width going to $\infty$, and both of these intervals contain exactly one change point (as their widths are $O(w(N)) = o(\delta_N)$). Hence, with probability $\to 1$ the multiple change point problem has simplified to $\hat{J}$ single change point problems, justifying ISM6 where a stump model is fitted inside each of $S^{(2)}(\hat{\tau}_j^{(1)})$'s.

**Asymptotic distributions of the intelligent sampling based estimators:** Next, we present results on the joint asymptotic distribution of the intelligent sampling estimates. There are two types of results which can be derived: a set of results that can accommodate for the case where the number of change points $J$ grows with $N$, and a more specific result for when $J$ stays constant with $N$.

In the first more general setting, we present a set of probability bounds that relate the deviation of the $\hat{\tau}^{(2)}$'s to the $L$-type distributions introduced in Section 2.2.1. This result characterizes the asymptotic behavior enough to allow the construction

of confidence intervals around the change point estimates. In the second setting, a more conventional distributional convergence result, in terms of joint probability mass functions, can be formulated.

**General Case:** For all $\alpha \in (0,1)$, let $Q_\Delta(1-\alpha)$ be the $1-\alpha$ quantile of $\left|\arg\min_{t\in\mathbb{Z}} X_\Delta(t)\right|$. An ideal result would be of the form:

$$\mathbb{P}\left[\hat{J} = J;\ \left|\lambda_2\left(\tau_j, \hat{\tau}_j^{(2)}\right)\right| \le Q_{\Delta_j}(\sqrt[J]{1-\alpha}) \text{ for all } j = 1,\dots,J\right] \to 1-\alpha \tag{2.16}$$

where $\lambda_2 := \lambda_{2,N}$ is a distance function to account for the omission of the first stage subsampling points:

$$\lambda_2(a,b) := \begin{cases} \sum_{i=1}^{N} 1\left(a < i \le b\right) \cdot 1(i \ne k\lfloor N/N_1 \rfloor \text{ for any integer } k) & \text{if } a \le b \\ -\sum_{i=1}^{N} 1\left(b < i \le a\right) \cdot 1(i \ne k\lfloor N/N_1 \rfloor \text{ for any integer } k) & \text{otherwise} \end{cases} \tag{2.17}$$

However, as the distribution of $\arg\min_{t\in\mathbb{Z}} X_\Delta(t)$ is discrete, it is generally not possible to get the probabilities exactly equal to $\sqrt[J]{1-\alpha}$. Instead, we derive the following result:

**Theorem II.3.** *For any $\alpha \in (0,1)$, define:*

$$P_\Delta(1-\alpha) = \mathbb{P}\left[\left|\arg\min_{t\in\mathbb{Z}} X_\Delta(t)\right| \le Q_\Delta(1-\alpha)\right]. \tag{2.18}$$

*Suppose that the first stage estimates satisfy (2.14) with a $\rho_N$ such that $J\rho_N \to 0$. Then*

$$\begin{aligned} &\mathbb{P}\left[\hat{J} = J;\ \left|\lambda_2\left(\tau_j, \hat{\tau}_j^{(2)}\right)\right| \le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \text{ for all } j = 1\dots,J\right] \\ = &\left(\prod_{j=1}^{J} P_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right) + o(1) \end{aligned} \tag{2.19}$$

$$\mathbb{P}\left[\hat{J} = J;\ \left|\lambda_2\left(\tau_j, \hat{\tau}_j^{(2)}\right)\right| \leq Q_{(|\Delta_j|+2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \text{ for all } j = 1\ldots, J\right]$$

$$= \left(\prod_{j=1}^{J} P_{(|\Delta_j|+2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right) + o(1) \tag{2.20}$$

*Proof.* See Section A.2.1 of Supplement Part B. □

The additional condition requiring $J\rho_N \to 0$ is due to the details of the proof, but this can be satisfied by existing change point methods, including the one that will be showcased later on in this paper. The proof also involves comparisons between random walks which appear during estimation, to the random walks $X_{(|\Delta_j|\pm 2\rho_N)/\sigma}(\cdot)$, and thus the results are in terms of the SNR's $(|\Delta_j| \pm 2\rho_N)/\sigma$. It is still possible to make sense of this in terms of the parameters $\Delta_j/\sigma$: due to the stochastic ordering of the $L_\Delta$ random variables,

$$Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \geq Q_{|\Delta_j|/\sigma}(\sqrt[J]{1-\alpha}) \geq Q_{(|\Delta_j|+2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}), \tag{2.21}$$

and therefore,

$$\mathbb{P}\left[\hat{J} = J;\ \left|\lambda_2\left(\tau_j, \hat{\tau}_j^{(2)}\right)\right| \leq Q_{|\Delta_j|/\sigma}(\sqrt[J]{1-\alpha}) \text{ for all } j = 1\ldots, J\right] \tag{2.22}$$

will be between the values (2.19) and (2.20) up to a $o(1)$ term. This shows that (2.22) is sandwiched between two sequences that have a liminf at least the value of $1-\alpha$, justifying the construction of confidence intervals using the $L_{\Delta_j/\sigma}$ distributions.

A second ramification of Theorem II.3 is the rate of convergence.

**Theorem II.4.** *Suppose conditions (M1) to (M4) are satisfied and the first stage estimates satisfy the consistency result (2.14). Then, for any $\varepsilon > 0$, there exist constants $C_1$ and $C_2$ (depending on $\epsilon$) such that*

$$\mathbb{P}\left[\hat{J} = J, \max_{k=1,\ldots,J} |\hat{\tau}_k^{(2)} - \tau_k| \leq C_1 \log(J) + C_2\right] \geq 1 - \varepsilon \tag{2.23}$$

*for all sufficiently large $N$.*

*Proof.* Fix any $\alpha \in (0,1)$. It is possible to show that for all large $N$, $|\hat{\tau}_j^{(2)} - \tau_j| \leq 2 * \lambda_2\left(\tau_j, \hat{\tau}_j^{(2)}\right)$, and using Lemma 8 of Supplement Part C, it is also possible to show that for all $j = 1, \ldots, J$,

$$Q_{|\Delta_j|/\sigma}(\sqrt[J]{1-\alpha}) \leq Q_{\underline{\Delta}/\sigma}(\sqrt[J]{(1-\alpha)}) \leq C_1 \log(C_2 J) \tag{2.24}$$

for some positive constants $C_1, C_2$ not changing with $N$ or $j$. These combine to give

$$\begin{aligned} & \mathbb{P}\left[\hat{J} = J;\ \frac{1}{2}\left|\hat{\tau}_j^{(2)} - \tau_j^{(2)}\right| \leq C_2 \log(C_2 J) \text{ for all } j = 1 \ldots, J\right] \\ \leq\ & \mathbb{P}\left[\hat{J} = J;\ \left|\lambda_2\left(\tau_j, \hat{\tau}_j^{(2)}\right)\right| \leq Q_{|\Delta_j|/\sigma}(\sqrt[J]{1-\alpha}) \text{ for all } j = 1 \ldots, J\right] \end{aligned} \tag{2.25}$$

We know from the previous discussion that, we have

$$\liminf_{N\to\infty} \mathbb{P}\left[\hat{J} = J;\ \left|\lambda_2\left(\tau_j, \hat{\tau}_j^{(2)}\right)\right| \leq Q_{|\Delta_j|/\sigma}(\sqrt[J]{1-\alpha}) \text{ for all } j = 1 \ldots, J\right] \geq 1 - \alpha \tag{2.26}$$

and therefore

$$\liminf_{N\to\infty} \mathbb{P}\left[\hat{J} = J;\ \frac{1}{2}\left|\hat{\tau}_j^{(2)} - \tau_j^{(2)}\right| \leq C_2 \log(C_2 J) \text{ for all } j = 1 \ldots, J\right] \geq 1 - \alpha \tag{2.27}$$

□

The conditions in which these two results (Theorems II.3 and II.4) hold include the possibility of letting the number of change points ($J$) grow to $\infty$. In the more specific case where $J$ stays finite as $N \to \infty$, both results would still hold, but an even more informative result can be derived.

**Finite $J$:** Consider the case where $J$ and the jump sizes stays constant with $N$. Specifically, suppose the following assumption is satisfied:

(M5): The number of change points $J$ is a finite constant not dependent on $N$, and the jump sizes $\Delta_j := \nu_j - \nu_{j-1}$ for $j = 1, ..., J$ are also constants not dependent on $N$.

Using this notation, we establish

**Theorem II.5.** *Suppose conditions (M1) to (M5), and the consistency condition (2.14) are satisfied. Then, the deviations* $\left\{\lambda_2\left(\tau_j, \hat{\tau}_j^{(2)}\right)\right\}_{j=1}^{\hat{J}}$ *jointly converge to the distribution of* $\left(L_{\frac{\Delta_1}{\sigma}}, ..., L_{\frac{\Delta_J}{\sigma}}\right)$*, where the* $L_{\frac{\Delta_i}{\sigma}}$*'s are mutually independent. That is, for any integers* $k_1, \ldots, k_J$*,*

$$\mathbb{P}\left[\hat{J} = J,\, \lambda_2\left(\tau_j, \hat{\tau}_j^{(2)}\right) = k_j \text{ for } 1 \le j \le J\right] \to \prod_{j=1}^{J} \mathbb{P}\left[L_{\Delta_j/\sigma} = k_j\right] . \tag{2.28}$$

*Proof.* See Section A.2.3 of Supplement Part B. □

**Remark 5.** *The theorems presented above are related: Theorem II.3 implies Theorem II.4, which in turn is used in the proof of Theorem II.5.* Alternatively*, one can demonstrate Theorem II.4* independently *of Theorem II.3, under more relaxed (sub-Gaussian) conditions on the error terms. Using this other method, one can arrive at a result very similar to Theorem II.5 where the* $\lambda_2(\tau_j, \hat{\tau}_j^{(2)})$ *converge to the argmin of the random walks defined in (2.4), with the* $\varepsilon_i^*$ *being iid and having identical distributions as the error terms. For details, the alternative proof of Theorem II.4 and its more relaxed assumptions are discussed in Section A.2.6 of the supplement, along with some comments regarding Theorem II.5 at the end of the section. Finally, we suspect that Theorem II.3 may also extend to broader classes of errors (beyond Gaussian) but may possibly entail the use of very different methods.*

## 2.4 Practical Implementation

In the previous section, we laid out a generic scheme for intelligent sampling which requires the use of a multiple change point estimation procedure on a sparse subsample of the *data-sequence*. Recall that any procedure that satisfies (2.13) can be used here. A variety of such procedures have been explored by various authors (such as *Venkatraman* (1992), *Fryzlewicz et al.* (2014), *Frick et al.* (2014), and *Bai and Perron* (1998)), and therefore a number of options are available. For the sake of concreteness, we pursue intelligent sampling with binary segmentation (henceforth abbreviated to 'BinSeg') employed at Step (ISM3). One main advantage of BinSeg is its computational scaling at an optimal rate of $O(N^* \log(N^*))$ when applied onto a data sequence of length $N^*$, and in addition it has the upside of being easy to program and popular in change point literature. We later discuss other potential options.

However, there are some issues involved in applying the results of BinSeg to our setting. First, BinSeg does not directly provide the signal estimators that are required in (2.13. We address this issue in Section 2.4.1, where we establish that given certain consistency conditions on the change points, which are satisfied by BinSeg, consistent signal estimators can be obtained by averaging the data between change point estimates. Second, there is no established method for constructing explicit confidence intervals for the actual change points using BinSeg, as existing results give orders of convergence but no asymptotic distributions or probability bounds with *explicit* constants. However, to implement intelligent sampling, one wants to have high-probability intervals around the initial change-point estimates on which to do the second round sampling, which requires calibration in terms of the coverage probability. To this end, in Section 2.4.2 we describe a procedure to be performed after applying BinSeg is applied on the first stage subsample: the extra steps provide us with explicit confidence intervals while not being slower than BinSeg in terms of order

of computational time.

Before we begin, we remind the reader that this section deals with the first stage subsample and not the whole sample. We will henceforth use the $\star$ notation in connection with the quantities involved at the first stage. So, we let $\nu_j^* := \nu_j$ for $j = 0, \ldots, J$ and $\rho_{N^*}^* := \rho_N$, and referring back to notation used in step (ISM2) and (ISM3), we consider the sub-dataset $Z_1, \ldots, Z_{N^*}$ as a multiple change point model, with change points $\tau_j^*$'s and levels $\nu_j^*$'s, following conditions (M1) to (M4) for all large $N$ (as a consequence of $Y_1, \ldots, Y_N$ satisfying conditions (M1) to (M4)). Using this notation, (2.13) translates to the requirement that a change point estimation scheme applied upon $Z_1, \ldots, Z_{N^*}$ w procures estimates $\hat{\tau}_j^*$'s and $\hat{\nu}_j^*$'s (equal to $\hat{\nu}_j^{(1)}$'s in (ISM3)) such that

$$\mathbb{P}\left[\hat{J} = J,\ \max_{j=1,\ldots J}\left|\hat{\tau}_j^* - \tau_j^*\right| \leq w^*(N^*),\ \max_{j=0,\ldots,J}\left|\hat{\nu}_j^* - \nu_j^*\right| \leq \rho_{N^*}^*\right] \to 1 \tag{2.29}$$

for some sequences $w^*(N^*)$ and $\rho_{N^*}^*$ such that $w^*(N^*) \to \infty$, $w^*(N^*) = o(\delta_{N^*}^*)$, and $\rho_{N^*}^* \to 0$ as $N^* \to \infty$. We, subsequently, refer to this latter condition.

### 2.4.1 Binary Segmentation

We first, briefly, describe the BinSeg algorithm(for a comprehensive exposition see *Fryzlewicz et al.* (2014)), where the length of the data sequence is denoted by $T$) and some additional properties that relate to our procedure. Consider the model given in (2.11). For any positive integers $1 \leq s \leq b < e \leq N^*$, let $n = e - s + 1$ and define the Cumulative Sum (CUSUM) statistic at $b$ with endpoints $(s, e)$ as

$$\bar{Z}_{s,e}^b = \sqrt{\frac{e-b}{n(b-s+1)}}\sum_{t=s}^{b} Z_t - \sqrt{\frac{b-s+1}{n(e-b)}}\sum_{t=b+1}^{e} Z_t.$$

Binary segmentation is performed by iteratively maximizing the CUSUM statistics over the segment between change point estimates, accepting a new change point if the maximum passes a threshold parameter $\zeta_{N^*}$. Specifically,

1. Fix a threshold value $\zeta_{N^*}$ and initialize the segment set $SS = \{(1, N^*)\}$ and the change point estimate set $\hat{\underline{\tau}}^* = \emptyset$.

2. Pick any ordered pair $(s, e) \in SS$, remove it from $SS$ (update $SS$ by $SS \leftarrow SS - \{(s, e)\}$). If $s \geq e$ then skip to step 5, otherwise continue to step 3.

3. Find the argmax and max of the CUSUM statistic over the chosen $(s, e)$ from the previous step: $b_0 = \underset{b \in \{s,\ldots,e-1\}}{\arg\max} |\bar{Z}^b_{s,e}|$ and $|\bar{Z}^{b_0}_{s,e}|$

4. If $|\bar{Z}^{b_0}_{s,e}| \geq \zeta_{N^*}$, then add $b_0$ to the list of change point estimates (add $b_0$ to $\hat{\underline{\tau}}^*$), and add ordered pairs $(s, b_0)$ and $(b_0 + 1, e)$ to $SS$, otherwise skip to step 5.

5. Repeat steps 2-4 until $SS$ contains no elements.

**Remark 6.** *For our model, this algorithm provides consistent estimates of* both *the location of the change points and the corresponding levels, given further restrictions on the minimal separation. Specifically, consistency results for BinSeg are limited to situations where the minimal separation distance between change points grows faster than the length of the data sequence taken to an appropriate power. Theorem 3.1 of Fryzlewicz et al. (2014) presents a concrete result in this direction, where this particular power* $\Theta$ *in their notation (which is* $1 - \Xi$ *in our notation) is restricted to be strictly larger than* $3/4$*. However, there is a caveat as far as this theorem is concerned. In similar later work (by the same author) on BinSeg in high dimensional settings, see Cho and Fryzlewicz (2015), the same spacing condition appears as Assumption (A1), but it turns out, on the basis of a corrigendum released by the authors Cho and Frylewicz that this spacing condition does not ensure consistency of BinSeg in that*

*paper; rather a stronger spacing condition,* $\Theta > 6/7$*, is needed. From our correspondence with the authors, there is strong reason to believe that the* $3/4$ *in Fryzlewicz et al. (2014)* should also change to $6/7$*, and accordingly, in the sequel where we focus on a BinSeg based approach, we restrict ourselves to this more stringent regime, to be conservative. An alternative is to use wild binary segmentation at Stage 1, which allows the separation between change-points to be of smaller order than BinSeg (therefore allowing more relaxed regimes), but suffers from the downside of increased computation time. Details are available in the supplement.*

We place the following two additional conditions:

(M6 (BinSeg)): $\Xi$ (from condition (M3)) is further restricted by $\Xi \in [0, 1/7)$,

(M7 (BinSeg)): $N_1$, from step (ISM2), is chosen so that $N_1 = K_1 N^\gamma$ for some $K_1 > 0$ and $\gamma > 7\Xi$.

The first of the above condition allows BinSeg to be consistent on some subsample of the data sequence, for if the condition was not satisfied and the minimum spacing $\delta_N$ grows slower than $N^{6/7}$, then established results on BinSeg (see Theorem II.6) could not guarantee consistency on any subsample of the data sequence (indeed, it won't guarantee consistency even BinSeg was applied on the entire data sequence). The latter of the above conditions means we choose a subsample large enough in order to use Theorem II.6, as this result would not work if the subsample is too small. When (M1) to (M6) are satisfied, the first stage subsample would have size $N^* = (K_1 + o(1))N^\gamma$ with minimal change point separation of $\delta^*_{N^*} = (N_1/N + o(1))\delta_N = (C + o(1))(N^*)^{1-\Xi/\gamma}$ for some positive constant $C$. This would allow us to apply the following BinSeg result:

**Theorem II.6.** *Suppose that conditions (M1) to (M4), (M6 (BinSeg)), and (M7 (BinSeg)) are satisfied, and the tuning parameter* $\zeta_{N^\star}$ *is chosen appropriately so that*

- *if* $\Xi/\gamma > 0$ *then* $\zeta_{N^*} = c_1(N^*)^{\xi}$ *where* $\xi \in (\Xi/\gamma, 1/2 - \Xi/\gamma)$ *and* $c_1 > 0$
- *if* $\Xi/\gamma = 0$ *then* $c_2 \left(\log(N^*)\right)^p \leq \zeta_{N^*} \leq c_3(N^*)^{\xi}$ *where* $p > 1/2, \xi < 1/2$, *and* $c_2, c_3 > 0$ .

*Define* $E_{N^*} = \left(\frac{N^*}{\delta^*_{N^*}}\right)^2 \log(N^*)$. *Then, there exist positive constants* $C, C_1$ *such that*

$$\mathbb{P}\Big[\hat{J} = J; \quad \max_{i=1,\ldots,J} |\hat{\tau}^*_i - \tau^*_i| \leq CE_{N^*}\Big] \geq 1 - C_1/N^* \,. \tag{2.30}$$

**Remark 7.** *The above theorem is adapted from Theorem 3.1 of Fryzlewicz et al. (2014) which applies to the more general setting where* $\underline{\Delta}$*, the minimum signal jump, can decrease to 0 as* $N \to \infty$*. As mentioned before, there is no easy recipe for determining* $C$ *explicitly, which is taken care of in the next section.*

We now propose estimators for the signals $\hat{\nu}^\star_j := \mathbb{E}[Z_{\tau^\star_j + 1}]$, for $j = 0, \ldots, J$. Intuitively, they can be estimated by the average of datapoints between each signal estimates:

$$\hat{\nu}^*_j = \frac{1}{\hat{\tau}^*_{j+1} - \hat{\tau}^*_j} \left( \sum_{\hat{\tau}^*_j < i \leq \hat{\tau}^*_{j+1}} Z_i \right) \qquad \text{for } j = 0, ..., \hat{J} \tag{2.31}$$

with the convention of $\hat{\tau}^*_0 := 0$ and $\hat{\tau}^*_{\hat{J}+1} := N^*$. These estimators are consistent:

**Lemma 1.** *Suppose conditions (M1) to (M4), (M6 (BinSeg)), and (M7 (BinSeg)) are satisfied, the* $\hat{\tau}^*_i$*'s are the BinSeg estimators, and* $\hat{\nu}^*_i$*'s are the signal estimators defined in (2.31). Then there exists a sequence* $\rho^*_{N^*} \to 0$ *such that* $J\rho^*_{N^*} \to 0$ *and*

$$\mathbb{P}\left[\hat{J} = J; \quad \max_{i=1,\ldots,J} |\hat{\tau}^*_i - \tau^*_i| \leq CE_{N^*}; \quad \max_{i=0,\ldots,J} |\hat{\nu}^*_i - \nu^*_i| \leq \rho^*_{N^*}\right] \to 1 \tag{2.32}$$

*as* $N^* \to \infty$.

*Proof.* See Section A.2.4 in Supplement Part B. □

By setting $\rho^*_{N^*} = \rho_N$ and $CE_{N^*} = w^*(N^*)$, condition (2.13), will be satisfied for a $\rho_N$ satisfying $J\rho_N \to 0$, which meets all requirement of Theorems II.3.

### 2.4.2 Calibration of intervals used in Stage 2

Constructing confidence intervals based on Theorem II.6 would require putting a value on $CE_{N^*} = C(N^*/\delta^*_{N^*})^2 \log(N^*)$ from (2.30). An estimate of $\delta^*_{N^*}$ can be obtained from the minimum difference of consecutive $\hat{\tau}^*_j$'s, but an explicit expression for $C$ is unavailable, and the existing literature on binary segmentation does not appear to provide such an explicit expression. To address this issue, we now introduce a calibration method which allows the construction of confidence intervals with explicitly calculable width around the first stage estimates $\hat{\tau}^{(1)}_j$'s: the idea is to fit stump models on data with indices $[\hat{\tau}^{(1)}_{j-1} + 1, \hat{\tau}^{(1)}_{j+1}]$, as each of these stretches forms a stump model with probability going to 1.

Consider starting from after step (ISM4) (e.g., Figure 2.1) where we have rough estimates $\hat{\tau}^*_i$'s of the change points (with respect to the $\{Z_i\}$ sequence) and $\hat{\nu}^{(1)}_i$'s of the signals, obtained from the $N^*$ sized subsample $\{Z_i\}$.

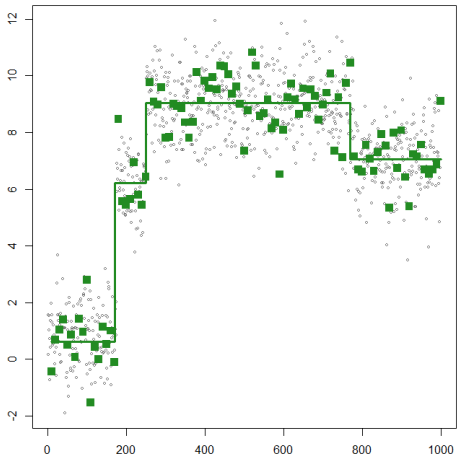


Figure 2.1: Green points are $Z_i$'s, solid green line is the BinSeg estimate.

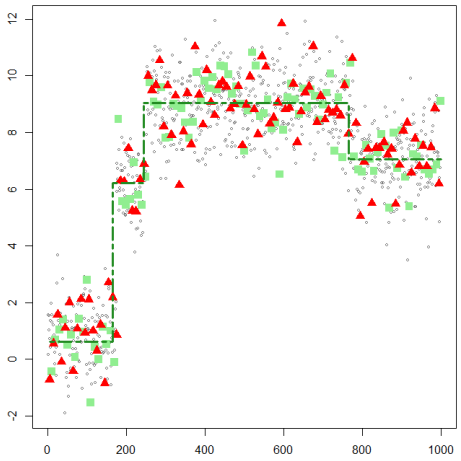


Figure 2.2: $Z_i$'s are light green points, BinSeg estimates as dashed green line, $V_i$'s as red points.

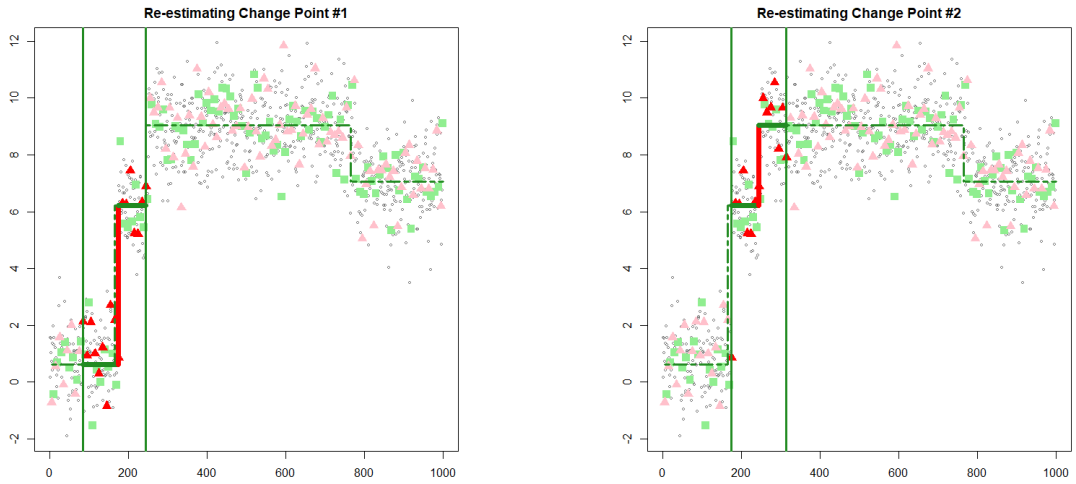


Figure 2.3: Re-estimation of the first (**Left Panel**) and second (**Right Panel**) detected change points (similar procedure for third estimated change point not shown). Solid green and solid red lines denote stump estimates using $V_i$'s from $\{V_k : \hat{\tau}_j^* - \hat{d}_j \leq k \leq \hat{\tau}_j^* + \hat{d}_j\}$ intervals.

We then pick a different subsample $\{V_i\}$ of equal size to the $\{Z_i\}$ subsample and consider the $\hat{\tau}_i^*$'s and $\hat{\nu}_i^{(1)}$'s as estimates for the parameters of this data sequence (e.g. Figure 2.2). For each $j$, fit a one-parameter stump model $f_j^{(d)}(k) = \hat{\nu}_{j-1}^{(1)} 1(k \leq d) + \hat{\nu}_j^{(1)} 1(k > d)$ (here $d$ is the discontinuity parameter) to the subset of $\{V_k : \hat{\tau}_{j-1}^* + 1 \leq k \leq \hat{\tau}_{j+1}^*\}$ given by $\{V_k : \hat{\tau}_j^* - \hat{d}_j \leq k \leq \hat{\tau}_j^* + \hat{d}_j\}$ where $\hat{d}_j = \min\{(\hat{\tau}_j^* - \hat{\tau}_{j-1}^*), (\hat{\tau}_{j+1}^* - \hat{\tau}_j^*)\}$, to get an updated least squares estimate of $\tau_j$ (e.g. Figure 2.3)[2].

More formally, the calibration steps are:

(ISM4-1): pick a positive integer $k_N$ less than $\lfloor N/N_1 \rfloor$ [3], take a subsample $\{V_i\}$ from the dataset of $\{Y_i\}$ which is the same size as the $\{Z_i\}$ subsample, by letting $V_i = Y_{i\lfloor N/N_1 \rfloor - k_N}$ for all $i$.

[2]We use this subset instead of the full interval to avoid situations where $\frac{\tau_j - \tau_{j-1}}{\tau_{j+1} - \tau_j} \to 0$ or $\infty$, which makes matters easier for theoretical derivations.

[3]A good pick is $k_N = \left\lfloor \frac{\lfloor N/N_1 \rfloor}{2} \right\rfloor$

The $\{V_i\}$ subsample also conforms to the model given in (2.11), with change points $\tau_i^{**} = \max\{j \in \mathbb{N} : j\lfloor N/N_1 \rfloor - k_N \leq \tau_i\}$ and minimum spacing $\delta_{N^*}^{**} := \min_k(\tau_{k+1}^{**} - \tau_k^{**})$ which satisfies $|\delta_{N^*}^{**} - \delta_{N^*}^{*}| \leq 1$.

(ISM4-2): For each $i = 1, ..., J$, consider the estimates $\hat{\tau}_i^*$ (obtained from the $\{Z_j\}$ subsample at step (ISM3)) as estimators for $\tau_i^{**}$. From (2.13) it is possible to derive that

$$\mathbb{P}\left[\hat{J} = J, \max_{i=1,...,J} |\hat{\tau}_i^* - \tau_i^{**}| \leq w^*(N^*) + 1, \max_{i=0,...,J} |\hat{\nu}_i^{(1)} - \nu_i| \leq \rho_N\right] \to 1 \quad (2.33)$$

where $w^*(N) + 1 \to \infty$ and $(w^*(N^*) + 1)/\delta_{N^*}^{**} \to 0$.

(ISM4-3): For each $i = 1, ..., \hat{J}$, define $\hat{d}_i = \min\{(\hat{\tau}_{i+1}^* - \hat{\tau}_i^*), (\hat{\tau}_i^* - \hat{\tau}_{i-1}^*)\}$ (with $\hat{\tau}_0^* = 0$ and $\hat{\tau}_{\hat{J}+1}^* = \lfloor N/\lfloor N/N_1 \rfloor \rfloor$), and re-estimate the change points by letting

$$\hat{\tau}_i^{re} := \underset{t:|t-\hat{\tau}_i^*|<\hat{d}_i}{\arg\min} \left[ \sum_{j:|j-\hat{\tau}_i^*|<\hat{d}_i} \left(V_i - \hat{\nu}_{i-1}^{(1)} 1(j \leq t) - \hat{\nu}_i^{(1)} 1(j > t)\right)^2 \right] \quad (2.34)$$

for $i = 1, ..., \hat{J}$.

(ISM4-4): To translate the $\hat{\tau}_i^{re}$'s (change point estimates for the subsample for the $\{V_j\}$'s) into estimates for $\tau_1, \ldots, \tau_J$ (change points for the full data set), set the first stage change point estimators as $\hat{\tau}_i^{(1)} := \hat{\tau}_i^{re}\lfloor N/N_1 \rfloor - k_N$.

**Remark 8.** *It is important to note that although the above steps are presented in the context of using BinSeg at first stage, in practice they can be used in many other situations. As far as intelligent sampling is concerned, any other change point estimation method which satisfies similar consistency conditions to BinSeg can be used in the first stage. More broadly, these steps can be used outside of the intelligent sampling framework. Given a consistent change point estimation scheme* for which a method to construct explicit confidence intervals is not known*, one could split the data in two*

*subsamples, the odd points (first, third, fifth, etc data points) and the even points. The aforementioned estimation scheme could be applied to the odd points, and afterwards steps (ISM4-1) to (ISM4-4) could be applied to the even points. A result similar to that of Theorem II.7, presented below, could then be used to construct confidence intervals.*

**Theorem II.7.** *Suppose conditions (M1) to (M4) are satisfied, and the estimation method used in step (ISM3) satisfies (2.13), and the pertinent $\rho_N$ appearing in (2.13) also satisfies $J\rho_N \to 0$. For any sequence $\alpha_N$ between 0 and 1 such that $\alpha_N \geq CN^{-\eta}$ for some positive $C$ and $\eta$, we have*

$$\mathbb{P}\left[\hat{J} = J, |\hat{\tau}_j^{re} - \tau_j^{**}| \leq Q_{|\Delta_j|/\sigma}\left(1 - \frac{\alpha_N}{J}\right) \text{ for } j = 1, \ldots, J\right] \geq 1 - \alpha_N + o(1) \tag{2.35}$$

*Proof.* See Section A.2.2 of Supplement Part B. □

**Remark 9.** *Similar to Theorem II.3, the condition $J\rho_N \to 0$ is required here. This is because the proofs of both results are similar in structure. For intelligent sampling with BinSeg at stage 1, we re-iterate that $J\rho_N \to 0$ is automatically satisfied, and we additionally remark that for such a procedure the $o(1)$ term in (2.35) is also $o(\alpha_N)$.*

The practical implication of this result for implementing intelligent sampling is that we can obtain explicitly calculable simultaneous confidence intervals for the change-points. The intervals $[\hat{\tau}_j^{re} \pm Q_{|\Delta_j|/\sigma}\left(1 - \frac{\alpha_N}{J}\right)]$ for $j = 1, \cdots, J$ capture the sparse scale change points $\tau_j^{**}$ for $j = 1, \cdots, J$ with probability approaching 1 if we choose some $\alpha_N \to 0$. Converting back to the original scale, the intervals $[\hat{\tau}_j^{(1)} \pm (Q_j\left(1 - \frac{\alpha_N}{J}\right) + 1)\lfloor N/N_1 \rfloor]$ for $j = 1, \cdots, \hat{J}$ have the properties that they simultaneously capture $\tau_1, \cdots, \tau_J$ with probability approaching 1.[4] The second stage

[4] In practice, the $\Delta_j/\sigma$'s are estimated from data.

samples are then picked as points within each of these intervals that are not divisible by $\lfloor N/N_1 \rfloor$ (refer to steps (ISM5) and (ISM6)).

### 2.4.3 Computational Considerations

The order of computational time is higher and needs more work to compute than in the single change point case, owing to the fact that the procedure can involve a growing number of data intervals at the second stage. For the sake of our analysis, we make the simplifying assumption that $\delta_N/N^{1-\Xi} \to C_1$ and $J(N)/N^{\Lambda} \to C_2$ for some $\Lambda \in [0, \Xi]$ and some positive constants $C_1, C_2$. As a reminder, for intelligent sampling with BinSeg at stage 1, conditions (M6 (BinSeg)) and (M7 (BinSeg)) automatically impose the condition that $\Lambda \leq \Xi < 1/7$.

The BinSeg procedure, when applied to a data sequence of $n$ points, takes $O(N\log(N))$ time to compute (see *Fryzlewicz et al.* (2014)). Since first stage of intelligent sampling involves applying BinSeg to $O(N^{\gamma})$ points, it therefore takes $O(N^{\gamma}\log(N))$ time to obtain the first stage estimators. After the BinSeg estimates are obtained, we use the method described in Section 2.4.2 to upgrade them to ones whose asymptotic distributions are known, this subsequent step only involving least squares fitting upon $O(N^{\gamma})$ points and therefore requiring only $O(N^{\gamma})$ computational time, leaving the total time as $O(N^{\gamma}\log(N))$ up to this point.

From here on, we use Theorem II.7 and construct confidence intervals $\left[\hat{\tau}_j^{(1)} \pm \left(Q_{\Delta_j/\sigma}(\sqrt[\hat{J}]{1-\alpha}) + 1\right) \lfloor \frac{N}{N_1} \rfloor\right]$ for $j = 1, \ldots, \hat{J}$. Lemma 8 (see Section A.3.2, Supplement Part C) tells us that $Q_{\Delta_j/\sigma}(\sqrt[\hat{J}]{1-\alpha})$ can be bounded by a multiple of $\log(\hat{J})$, and therefore conditional on the value of $\hat{J}$, the second stage of intelligent sampling will involve least squares fitting on $O\left(\hat{J}N^{1-\gamma}\log(\hat{J})\right)$ points, taking $O\left(\hat{J}N^{1-\gamma}\log(\hat{J})\right)$ time to compute. Although the distribution of the $\hat{J}$ obtained

from BinSeg is not fully known, a consequence of Theorem II.6 is that $\mathbb{P}\left[\hat{J} = J\right] \geq 1 - CN^{-1}$ for some constant $C$, and therefore

$$\mathbb{E}\left[\hat{J}\log(\hat{J})\right] \leq J\log(J) + C\frac{N\log(N)}{N} = O(J\log(N)). \tag{2.36}$$

This leads to the conclusion that the second stage has a computational time that is on average $O\left(JN^{1-\gamma}\log(N)\right) = O\left(N^{1-\gamma+\Lambda}\log(N)\right)$, and the entire procedure takes $O\left(N^{\gamma\vee(1-\gamma+\Lambda)}\log(N)\right)$ time.

Using this result we could choose an optimal $\gamma$ and obtain the optimal computational time for each value of $\Xi \in [0, 1/7)$ and $\Lambda \in [0, \Xi]$. This can be done by setting the order of the first stage ($O(N^{\gamma}\log(N))$) to equal the order of time for the second stage ($O(N^{1-\gamma+\Lambda}\log(N))$) which would be $\gamma = \frac{1+\Lambda}{2}$. However Condition (M7 (BinSeg)) prevents this from being done everywhere by placing the restriction that $\gamma > 7\Xi$. Thus $\gamma_{min}$ would be the maximum of $\frac{1+\Lambda}{2}$ and $7\Xi + \eta$ ($\eta$ any tiny positive value), resulting in order $N^{\gamma_{min}\vee(1-\gamma_{min}+\Lambda)}\log(N)$ computational time.

- For $\Xi \in [0, 1/14)$, we have $\frac{1+\Lambda}{2} < 4\Xi$ and hence $\gamma_{min} = \frac{1+\Lambda}{2}$ and the computational time is order $N^{(1+\Lambda)/2}\log(N)$.
- For $\Xi \in [1/13, 1/7)$, we have $\frac{1+\Lambda}{2} > 7\Xi$, hence $\gamma_{min} = \frac{1+\Lambda}{2}$ and the computational time is order $N^{4\Xi+\eta}\log(N)$.
- For $\Xi \in [1/14, 1/13)$, $\gamma_{min}$ can be either $\frac{1+\Lambda}{2}$ or $4\Xi + \eta$, whichever is greater, and the computational time would be either $N^{(1+\Lambda)/2}\log(N)$ or $N^{7\Xi+\eta}\log(N)$. respectively.

Table 2.1: Table of $\gamma_{min}$ and computational times for various values of $\Xi$. Also shown are their values for extreme value of $\Lambda$ ($\Lambda = 0$ and $\Lambda = \Xi$). For $\Xi \geq 1/7$ no values of $\gamma$ will allow us to obtain consistency from Theorem II.6

| $\Xi$ | $[0, 1/14)$ | $[1/14, 1/13)$ | [1/13,1/7) | [1/7,1] |
|---|---|---|---|---|
| $\gamma_{min}$ | $\frac{1+\Lambda}{2}$ | $\max\left\{\frac{1+\Lambda}{2}, 7\Xi+\eta\right\}$ | $7\Xi+\eta$ | N/A |
| Order of Time | $N^{(1+\Lambda)/2}\log(N)$ | $\max\{N^{(1+\Lambda)/2}, N^{7\Xi+\eta}\}\cdot\log(N)$ | $N^{7\Xi+\eta}\log(N)$ | N/A |
| $\gamma_{min}$ $(\Lambda=0)$ | $\frac{1}{2}$ | $7\Xi+\eta$ | $7\Xi+\eta$ | N/A |
| Time $(\Lambda=0)$ | $N^{1/2}\log(N)$ | $N^{7\Xi+\eta}\log(N)$ | $N^{7\Xi+\eta}\log(N)$ | N/A |
| $\gamma_{min}$ $(\Lambda=\Xi)$ | $\frac{1+\Xi}{2}$ | $\frac{1+\Xi}{2}$ | $7\Xi+\eta$ | N/A |
| Time $(\Lambda=\Xi)$ | $N^{(1+\Xi)/2}\log(N)$ | $N^{(1+\Xi)/2}\log(N)$ | $N^{7\Xi+\eta}\log(N)$ | N/A |

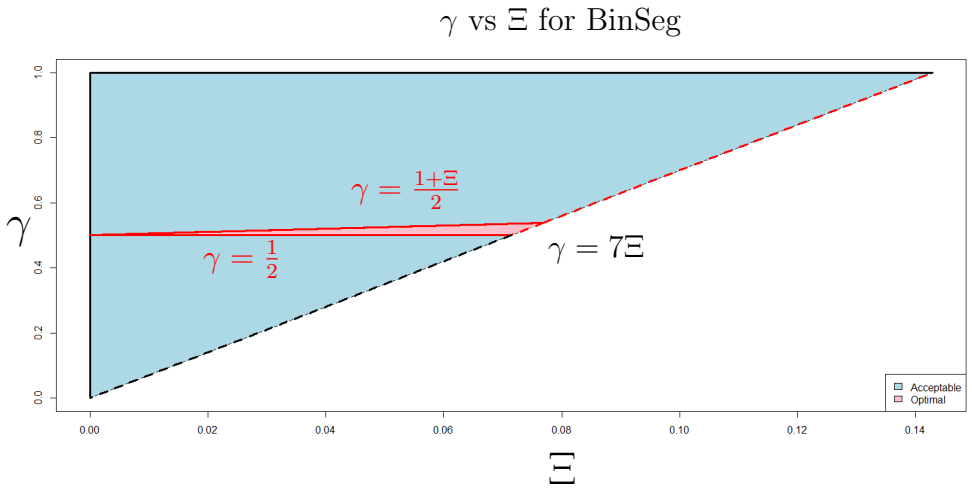


Figure 2.4: Blue triangle encompasses all valid values of $\gamma$ vs $\Xi$ as set by (M7 (BinSeg)). Pink region, solid red lines, and dotted red lines denotes $\gamma_{min}$ for each $\Xi$ ($\gamma_{min}$ can vary for different values of $\Lambda$ even when $\Xi$ is fixed, hence the red region).

It can be seen that the biggest decrease in order of average computational time is for small values of $\Xi$ and $\Lambda$, and in fact for $\Xi < 1/14$ and $\Lambda = 0$ it is $O(\sqrt{N}\log(N))$, which is marginally slower than intelligent sampling on a single change point. For larger values of $\Xi$, there is less than a square root drop in $N\log(N)$ (order of using BinSeg on the whole data) to $N^{\gamma_{min}}\log(N)$ (intelligent sampling), to the point where as $\Xi \to 1/7$, both procedures take near the same order of time.

**Remark 10.** *Note that when implementing the intelligent sampling strategy knowl-*

*edge of* $\Xi$ *is desirable, but in practice, its value is unknown. If one is willing to impose an upper bound on* $\Xi$*, intelligent sampling can be implemented with this (conservative) upper-bound.*

**Remark 11. Multistage Intelligent Sampling in the multiple change-point problem:** *We can also consider intelligent sampling with multiple* ($> 2$) *stages of estimation for model (2.11). An* $m$*-stage intelligent sampling procedure would entail:*
**a.** *Take a uniform subsample* $Y_{\lfloor N/N_1 \rfloor}, Y_{2\lfloor N/N_1 \rfloor}, Y_{3\lfloor N/N_1 \rfloor}, \ldots,$ *where* $N_1 = KN^{\gamma}$ *for some* $K > 1, \gamma \in (0,1)$*, to obtain estimates* $\hat{J}$*,* $\hat{\tau}_1^{(1)}, \ldots, \hat{\tau}_{\hat{J}}^{(1)}$*, and confidence intervals* $\left[\hat{\tau}_j^{(1)} - w(N), \hat{\tau}_j^{(1)} + w(N)\right]$*,* $1 \leq j \leq \hat{J}$*, for the change points.*
**b.** *On each interval* $\left[\hat{\tau}_j^{(1)} - w(N), \hat{\tau}_j^{(1)} + w(N)\right]$ *for* $1 \leq j \leq \hat{J}$ *perform the* $(m-1)$ *stage intelligent sampling procedure for the single change point (as described in Remark 2).*

Computationally, an $m$ stage procedure works faster than a two stage procedure.

## 2.5 Sample Size Considerations from a Methodological Angle

In the asymptotic setting of Section 2.4.3, we were concerned about minimizing the *order* of computational time required for locating the change points through intelligent sampling, assuming that certain important quantities were known. The focus in this section is on obtaining explicit expressions for the minimum sample size that the procedure requires to correctly identify the underlying change points. Obviously, the minimum sample size is the key driver in the computational time formlas provided, albeit not the single one, and also addresses computer memory usage issues. In order to develop explicit expressions for the total computational time, one would need to know exactly how fast BinSeg runs versus data size, in terms of its model parameters[5] and this is unavailable as an exact expression. Therefore,

[5]For example, BinSeg will generally terminate in fewer steps on a dataset with fewer change points than on another dataset of the same length but more change points.

we look at minimizing the subsample utilized as a proxy, with the added benefit of deriving the least amount of data that must be held in memory at a single time.

We have already investigated the optimal order of the first stage subsample, denoted $N_1$, and demonstrated in Section 2.4.3 that in the best cases the size of both the first and second stage subsamples scales as $\sqrt{N}\log(N)$. Although valid, these previous analyses only apply to an abstract asymptotic setting. In practice, given a data set with fixed (large) $N$, a different approach is needed to determine the optimal number of points to use at each different stage.

Given the number of change points and their associated SNR's, we show below how to optimally allocate samples in order to minimize the total number used, for intelligent sampling. We start with the two-stage intelligent sampling procedure and assume that in stage 1, roughly $N_1$ points are used for BinSeg and another $N_1$ points, for the calibration steps described in Section 2.4.2. At stage 2, we work with $\hat{J}$ (which is $\approx J$) intervals. Using Theorem II.7, setting the width of the second stage intervals to be $\left(Q_\Delta\left(1-\frac{\alpha}{J}\right)+1\right)\left\lfloor\frac{N}{N_1}\right\rfloor$ for a small $\alpha$ will ensure that they cover the true change points with high probability (close to $1-\alpha$ if not greater). Assuming $N_1$ is large enough so that the first stage is accurate (ie $\hat{J}=J$ and $\max_j|\hat{\tau}_j^{(1)}-\tau_j|$ is small with high probability) , the number of points used in the two stages, combined, is approximately

$$2N_1+\frac{2\left(\sum_{j=1}^{J}\left(Q_{\Delta_j/\sigma}\left(1-\frac{\alpha}{J}\right)+1\right)\right)N}{N_1}. \tag{2.37}$$

This presents a trade-off, e.g. if we decrease $N_1$ by a factor of 2, the second term in (2.37) increases by a factor of 2. To use a minimal number of points in both stages, we need to set $N_1=\sqrt{N\sum_{j=1}^{J}\left(Q_{\Delta_j/\sigma}\left(1-\frac{\alpha}{J}\right)+1\right)}$. In turn this yields a minimum of $4\sqrt{N\sum_{j=1}^{J}\left(Q_{\Delta_j/\sigma}\left(1-\frac{\alpha}{J}\right)+1\right)}$ when plugged into (2.37). For any given values of $N$, $J$, and SNR, this provides a lower bound on the minimum number of points that

intelligent sampling must utilize, and Tables 2.5 and 2.6 depict some of these lower bounds for a select number of these parameters.

Figure 2.5: For $N = 1.5 \times 10^7$, the minimal percentage of data that must be used for various values of $J$ and SNR, assuming all jumps have equal SNR and $\alpha = 0.01$.

| $\Delta/\sigma$ \ $J$ | 40 | 60 | 80 | 100 |
|---|---|---|---|---|
| 1 | 4.33 | 5.43 | 6.4 | 7.3 |
| 1.25 | 3.46 | 4.38 | 5.14 | 5.84 |
| 1.5 | 2.92 | 3.67 | 4.33 | 4.84 |
| 1.75 | 2.53 | 3.2 | 3.7 | 4.26 |
| 2 | 2.26 | 2.77 | 3.33 | 3.72 |
| 2.25 | 1.96 | 2.53 | 2.92 | 3.43 |
| 2.5 | 1.85 | 2.26 | 2.61 | 3.1 |
| 2.75 | 1.73 | 2.12 | 2.44 | 2.73 |
| 3 | 1.6 | 1.96 | 2.26 | 2.53 |

Figure 2.6: For $N = 1.5 \times 10^{10}$, the minimal percentage of data that must be used for various values of $J$ and SNR, assuming all jumps have equal SNR and $\alpha = 0.01$.

| $\Delta/\sigma$ \ $J$ | 100 | 250 | 500 | 1000 |
|---|---|---|---|---|
| 1 | 0.231 | 0.386 | 0.57 | 0.839 |
| 1.25 | 0.185 | 0.31 | 0.456 | 0.669 |
| 1.5 | 0.157 | 0.258 | 0.386 | 0.566 |
| 1.75 | 0.135 | 0.225 | 0.327 | 0.484 |
| 2 | 0.118 | 0.2 | 0.292 | 0.426 |
| 2.25 | 0.108 | 0.179 | 0.263 | 0.386 |
| 2.5 | 0.098 | 0.163 | 0.231 | 0.343 |
| 2.75 | 0.086 | 0.146 | 0.219 | 0.31 |
| 3 | 0.08 | 0.137 | 0.207 | 0.292 |

Note that while the fraction of points used for the larger $N$ above is smaller, in absolute terms, this still translates to very large subsamples: even just 0.57% of $1.5 \times 10^{10}$ (for SNR 1 and $J = 500$ on Table 2.6) is a very large dataset of $8.6 \times 10^7$, which almost requires server type computer capabilities. The situation becomes more tenuous for larger values of $N$. This suggests that *a larger number of stages* is in order for sample sizes of $N$ exceeding $10^{10}$.

For a three-stage implementation, suppose $\approx N_1$ points are utilized at stage 1, letting

us form simultaneous confidence intervals that are (approximately) of the form

$$\left[\hat{\tau}_j^{(1)} - \sum_{j=1}^{J}\left(Q_{\Delta_j/\sigma}\left(1-\frac{\alpha}{J}\right)+1\right)\frac{N}{N_1}, \hat{\tau}_j^{(1)} + \sum_{j=1}^{J}\left(Q_{\Delta_j/\sigma}\left(1-\frac{\alpha}{J}\right)+1\right)\frac{N}{N_1}\right] \quad (2.38)$$

$$\text{for } j = 1, \ldots, J \quad (2.39)$$

(assuming $\hat{J} = J$ for simplification). At stage 2, suppose at the $j$'th confidence interval we subsample roughly $N_2^{(j)}$ points, giving us a subsample which skips approximately every $2Q_{\Delta_j/\sigma}\left(1-\frac{\alpha}{J}\right)\frac{N}{N_1 N_2^{(j)}}$ points. Hence, at stage 3 we work with confidence intervals that are (approximately) of the form

$$\left[\hat{\tau}_j^{(2)} \pm \left(Q_{\Delta_j/\sigma}\left(1-\frac{\alpha}{J}\right)+1\right)\left(2\left(Q_{\Delta_j/\sigma}\left(1-\frac{\alpha}{J}\right)+1\right)\frac{N}{N_1 N_2^{(j)}}\right)\right] \quad (2.40)$$

for $j = 1, \ldots, J$. In total all three stages use around a total of

$$2N_1 + \sum_{j=1}^{J} N_2^{(j)} + \frac{4N}{N_1}\left(\sum_{j=1}^{J}\frac{\left(Q_{\Delta_j/\sigma}\left(1-\frac{\alpha}{J}\right)+1\right)^2}{N_2^{(j)}}\right) \quad (2.41)$$

points. This expression is minimized by setting

$$N_1 = N^{1/3}\left(\sum_{k=1}^{J}\left(Q_{\Delta_j/\sigma}\left(1-\frac{\alpha}{J}\right)+1\right)\right)^{2/3} \text{ and}$$
$$N_2^{(j)} = 2N^{1/3}\frac{Q_{\Delta_j/\sigma}\left(1-\frac{\alpha}{J}\right)+1}{\left(\sum_{k=1}^{J}\left(Q_k(\alpha, J)+1\right)\right)^{1/3}} \quad (2.42)$$

for $j = 1, \ldots, J$, which in turn gives a minimum of $6N^{1/3}\left(\sum_1^J\left(Q_{\Delta_j/\sigma}\left(1-\frac{\alpha}{J}\right)+1\right)\right)^{2/3}$ for (2.41). A similar analysis on a four-stage procedure would have the optimal subsample allocation as $N_1 = N^{1/4}\left(\sum_{k=1}^{J}\left(Q_{\Delta_j/\sigma}\left(1-\frac{\alpha}{J}\right)+1\right)\right)^{3/4}$ and $N_2^{(j)} = N_3^{(j)} = 2\left(Q_{\Delta_j/\sigma}\left(1-\frac{\alpha}{J}\right)+1\right)N^{1/4}\left(\sum_{k=1}^{J}\left(Q_k(\alpha, J)+1\right)\right)^{-1/4}$ for $j = 1, \ldots, J$, which yields a total of $8N^{1/4}\left(\sum_{k=1}^{J}\left(Q_{\Delta_j/\sigma}\left(1-\frac{\alpha}{J}\right)+1\right)\right)^{3/4}$ points utilized.

Figure 2.7: For $N = 1.5 \times 10^{10}$, minimal percentage of the data that must be used for a three stage procedure, assuming all jumps have equal SNR and $\alpha = 0.01$.

| $\Delta/\sigma$ \ $J$ | 100 | 250 | 500 | 1000 |
|---|---|---|---|---|
| 1 | 0.03 | 0.061 | 0.101 | 0.168 |
| 1.25 | 0.023 | 0.045 | 0.075 | 0.127 |
| 1.5 | 0.018 | 0.036 | 0.06 | 0.099 |
| 1.75 | 0.015 | 0.029 | 0.049 | 0.082 |
| 2 | 0.012 | 0.025 | 0.041 | 0.068 |
| 2.25 | 0.011 | 0.022 | 0.036 | 0.06 |
| 2.5 | 0.009 | 0.018 | 0.031 | 0.052 |
| 2.75 | 0.009 | 0.017 | 0.027 | 0.046 |
| 3 | 0.008 | 0.016 | 0.025 | 0.043 |

Figure 2.8: For $N = 1.5 \times 10^{12}$, minimal percentage of the data that must be used for a four stage procedure, assuming all jumps have equal SNR and $\alpha = 0.01$.

| $\Delta/\sigma$ \ $J$ | 250 | 500 | 1000 | 2000 |
|---|---|---|---|---|
| 1 | 8.3e-04 | 1.5e-03 | 2.6e-03 | 4.6e-03 |
| 1.25 | 6.0e-04 | 1.1e-03 | 1.9e-03 | 3.3e-03 |
| 1.5 | 4.6e-04 | 8.2e-04 | 1.4e-03 | 2.5e-03 |
| 1.75 | 3.6e-04 | 6.6e-04 | 1.1e-03 | 2.0e-03 |
| 2 | 3.1e-04 | 5.5e-04 | 9.6e-04 | 1.7e-03 |
| 2.25 | 2.5e-04 | 4.5e-04 | 8.0e-04 | 1.4e-03 |
| 2.5 | 2.2e-04 | 4.0e-04 | 6.8e-04 | 1.2e-03 |
| 2.75 | 1.9e-04 | 3.5e-04 | 5.9e-04 | 1.1e-03 |
| 3 | 1.8e-04 | 3.0e-04 | 5.5e-04 | 9.2e-04 |

Comparing Figures 2.6 and 2.7, we focus on the case of 1000 change points with SNR 1.5: using three stages allows us to decrease the minimal required points by a factor of around five. The ease on computations is greater when looking at the largest amount of data the computer must handle at a time:[1] this is $N_1 \approx 2.1 \times 10^7$ for two stages and $N_1 \approx 2.5 \times 10^6$ for three stages, a decrease by a factor of 9. Meanwhile for a dataset of size 1.5 trillion, using four stages allows us to work with subsamples of size at most $N_1 \approx 4.7 \times 10^6$ for the more demanding scenario of SNR 1.5 and 2000 change points, a very manageable dataset for most computers.

We note here that these optimal allocations are valid assuming that BinSeg is able to pin down $\hat{J}$ and the change points with the initial subsample. In general, this will be the case provided the SNR is reasonable, and the initial subsample is large enough so that the change-points are adequately spaced apart. For example, in the

[1]For intelligent sampling the largest data subset the computer has to work with and hold in memory at any moment, under these optimal allocations and when all change points have equal SNR, is the roughly $N_1$ sized data set used at the initial step for BinSeg. All subsequent steps can work with sub-intervals of data less than $N_1$ in size.

context of the above tables, one can ask whether BinSeg will accurately estimate the parameters on a 2.4 million length dataset with 1000 evenly spaced change points, or 2000 change points on a 4.7 million length data with 2000 evenly spaced change points, under a constant SNR of 1.5 (which is of modest intensity). To this end, we ran a set of simulations and concluded that if there are over 1000 data points between consecutive change points of SNR 1.5, based on these two settings and for appropriate tuning parameters, BinSeg's estimators satisfy $\hat{J} = J$ and $\max |\hat{\tau}_j - \tau_j| \leq 150$ with probability over 99%.

Observe also that the formulas provided depend on the values of the SNRs at the change points and the actual number of change points ($J$). In practice, neither will be known, and the the practitioner will not be able to determine the derived allocations exactly. In such situations, conservative lower bounds on the SNRs and a conservative higher bound on $J$, can yield valid (but conservative) sampling allocations when plugged in to the expressions derived through this section. Such bounds can be obtained if background information about the problem and the data are available, or via rough pilot estimates on an appropriately sparse subsample.

It is also worth pointing out that the intelligent sampling procedure is readily adaptable to a distributed computing environment, which can come into play, especially with data sets of length exceeding $10^{12}$ that are stored sequentially across several storage disks. In such cases, the two sparse subsamples at the first stage, which are of much smaller order, can be transferred over to a central server (a much easier exercise than transferring all the data on to one server), where the first is analyzed via binary segmentation to determine the initial change-points, and the other used for the re-estimation procedure and associated confidence intervals as described in Section 2.4.2. As the number of disks on which the data are stored is of a much smaller order than the length of the data, each re-estimated change-point and its associated confidence interval will typically belong to a stretch of data completely contained

within one storage disk, and the subsequent resampling and estimation steps can be performed on the local processor, after the information on the confidence interval has been transferred back from the central server. An occasional communication between two machines may be necessary.

## 2.6 Dependent Errors

The proposed intelligent sampling procedure for multiple change point problems has so far been presented in the setting of i.i.d. data for a signal-plus-noise model. However, many data sequences (such as time series) usually exhibit temporal correlation. Hence, it is of interest to examine the properties of the procedure under a non-i.i.d. data generating mechanism.

While we believe that results akin to Theorem 3 (growing number of change-points in the i.i.d. error regime) should go through under various forms of dependence among errors, a theoretical treatment of this would require a full investigation of the tail properties of random walks under dependent increments and is outside the scope of this paper. An asymptotic distributional result, analogous to Theorem 5, under finitely many change points in the dependent regime is also expected to hold.

We present below a proposition for the finite $J$ case under a set of high-level assumptions.

Suppose that the data sequence is in the form (2.11) and satisfies conditions (M1) to (M3) and (M5). Upon the error terms, we impose the assumption that they have an autocorrelation structure which dies out at a polynomial rate or faster, and locally around the change points assume that the joint distributions of the errors are fixed [i.e. invariant to $N$]:

(M4-alt1): $\varepsilon_j$'s are each marginally $N(0, \sigma_j^2)$, and there exist positive constants $\sigma_{max}$, $B$ and $\alpha$, independent of $N$, such that $\sigma_j \leq \sigma_{max}$ and $\mathrm{cor}(\varepsilon_j, \varepsilon_{j+k}) \leq Bk^{-\alpha}$ for

any $j$ and $j+k$ from 1 to $N$.

(M4-alt2): there exists a sequence $w_e(N) \to \infty$ and Gaussian sequences $\{\epsilon_{i,j}\}_{i\in\mathbb{Z}}$ (not required to be stationary) for $j = 1, \ldots, J$, such that for all $j = 1, \ldots, J$ and all sufficiently large $N$, $\{\varepsilon_{\tau_j - w_e(N)}, \ldots, \varepsilon_{\tau_j + w_e(N)}\}$ has the same joint distribution as $\{\epsilon_{-w_e(N),j}, \ldots, \epsilon_{w_e(N),j}\}$.

On a set of data where (M4-alt1) and (M4-alt2) hold (along with assumptions (M1) to (M3), and (M5)), we want steps (ISM1) and (ISM4) to go through with some procedure that ensures

$$\mathbb{P}\left[\hat{J} = J,\ \max_{i=1,\ldots,J} |\hat{\tau}_i^{(1)} - \tau_i| \le w(N),\ \max_{i=0,\ldots,J} |\hat{\nu}_i^{(1)} - \nu_i| \le \rho_N\right] \to 1. \tag{2.43}$$

for some sequence $w(N) \to \infty$ and $w(N) = o(\delta_N)$.

Next, we desire for the final estimators $\hat{\tau}^{(2)}$ to be $O_p(1)$ consistent and have the property that for each $\epsilon > 0$ there exists a constant $C$ such that

$$\mathbb{P}\left[\hat{J} = J;\quad \max_{i=1,\ldots,J} \left|\hat{\tau}_i^{(2)} - \tau_i\right| \le C\right] \ge 1 - \epsilon\,. \tag{2.44}$$

for all sufficiently large $N$.

**Proposition 3.** *Suppose conditions (M1) to (M3), (M4-alt1), (M4-alt2), and (M5) are satisfied. Next, suppose the first stage estimators satisfy (2.43) and the second stage estimators, constructed as in the i.i.d. setting but with a minor modification*[2] *satisfy (2.44). Define the random walks*

$$Z_{i,j} = \begin{cases} \Delta_j(\epsilon_{1,j} + \cdots + \epsilon_{i,j}) - i\Delta_j^2/2, & i > 0 \\ 0, & i = 0 \\ \Delta_j(\epsilon_{i+1,j} + \cdots + \epsilon_{0,j}) - i\Delta_j^2/2, & i < 0\,, \end{cases} \tag{2.45}$$

[2]See the remark right after the proposition.

*with the $\epsilon_{i,j}$'s from condition (M4-alt2), for $j = 1, 2, \ldots, J$, and denote $\tilde{L}_j := \arg\min_{i \in \mathbb{Z}} Z_{i,j}$. Then $|\hat{\tau}_j^{(2)} - \tau_j|$'s for $j = 1, \ldots, J$ jointly converge to the distribution of $(\tilde{L}_1, ..., \tilde{L}_J)$: for any integers $k_1, \cdot, k_J$,*

$$\mathbb{P}\left[\hat{J} = J,\, |\hat{\tau}_j^{(2)} - \tau_j| = k_j \text{ for } 1 \leq j \leq J\right] \to \prod_{j=1}^{J} \mathbb{P}[\tilde{L}_j = k_j] \tag{2.46}$$

**Remark 12.** *As in the i.i.d. case, the intervals $[\hat{\tau}_j^{(1)} - Kw(N), \hat{\tau}_j^{(1)} + Kw(N)]$ for $j = 1, \ldots, J$ [obtained at step (ISM5)] would each contain only one change point with probability approaching one. We are therefore still justified in fitting stump models on each interval, although with a slight modification. Unlike the i.i.d. error terms scenario, the joint distribution of the error terms at the second stage does change when we condition on the estimators $\hat{\tau}_j^{(1)}$'s, regardless if we leave out the $\{Z_j\}$ subsample at the second stage. We thus make the following modification to (ISM5) in 2.3.1 (and assume this altered procedure is used from here on in this section):*

*(ISM5-alt): take $S^{(2)}\left(\hat{\tau}_i^{(1)}\right)$ as all integers in $[\hat{\tau}_i^{(1)} - Kw(N), \hat{\tau}_i^{(1)} + Kw(N)]$ without any points omitted.*

*Step (ISM-6) then proceeds as before.*

**Remark 13.** *We note that the asymptotic distribution given above and in Theorem II.5 have the same form, since as before, conditional on (2.43) being true, intelligent sampling simplifies the problem into multiple single change point problems. Using Proposition 3 to construct confidence intervals in a practical setting requires an idea of the joint distribution of the $\{\epsilon_{ij}\}$'s. In practice, one would have to impose some structural conditions, e.g. assuming an ARMA or ARIMA structure on long stretches of the errors to the left and right of the change points.*

**Remark 14.** *The hard work lies in the verification of the high-level conditions (2.43) and (2.44) in different dependent error settings, and as mentioned previously, is not*

*dealt with in this paper but should constitute interesting agenda for follow-up research on this problem. We will, however, use Proposition 3 in our simulation and data analysis sections to construct confidence intervals for various dependent error scenarios.*

## 2.7 Performance Evaluation of Intelligent Sampling Simulation Results

We next illustrate, through a set of simulation studies, the theoretical properties of the intelligent sampling procedure: the rate of convergence and the lower than $O(N\log(N))$ computational running time, along with the validity of the asymptotic distribution. All simulations in this section were performed on a server with 24 Intel Xeon ES-2620 2.00 GHz CPUs, with a total RAM of 64 GB.

**Implementation of Intelligent Sampling via BinSeg:** There are numerous parameters associated with the multiple change point problem. Of importance are not only just the minimal separation $\delta_N$, minimum jump size $\underline{\Delta}$, and the number of change points $J$, which are the main parameters that appear in the theory, but also how the change points are distributed across the data sequence (which can vary wildly if $(J+1)\delta_N < N$), the actual values of the jumps $\nu_{i+1} - \nu_i$, for $i = 0, ..., J$, and the first-stage subsample size $N_1$. All of these can affect the accuracy of the procedure, particularly in the first stage rather than the zoom in estimation at the second stage (which is usually accurate if the first stage was accurate to begin with).

In addition to the re-estimation procedure described in Section 2.4.2 , we also included some ad-hoc methods to practically improve the accuracy of binary segmentation for the sparse subsample. For the initial subsample $Z_1, Z_2, ...$ with binseg

estimates $\hat{\tau}_1^*, ..., \hat{\tau}_{\hat{J}}^*$ as was described in step (ISM2), consider the two drop steps: fix positive constants $\delta_D$ and $\Delta_D$, and,

(D1): if for some $1 < i \leq \hat{J}$ we have $|\hat{\tau}_i^* - \hat{\tau}_{i-1}^*| \leq \delta_D$, then remove $\hat{\tau}_i^*$ from the list of estimates;

(D2): continuing to denote the remaining estimates as $\hat{\tau}_1^*, ..., \hat{\tau}_{\hat{J}}^*$ for convenience, let each $\hat{\nu}_j^*$ be the mean of the $Z_i$'s from $\hat{\tau}_j^* + 1$ to $\hat{\tau}_{j+1}^*$; if for some $i$ we have $\left|\hat{\nu}_i^* - \hat{\nu}_{i-1}^*\right| \leq \Delta_D$ then drop $\hat{\tau}_i^*$ from the list of estimates.

The intuition behind the first step is to set $\delta_D$ as a reasonably small integer, then for any dataset where $\delta_N$ is large, no two change points should be $\delta_D$ apart. Similarly, when $\Delta_D$ is set to be a number lower than $\underline{\Delta}$ (or some estimate thereof), step (D2) drops any estimate which does not exhibit a large enough signal change before and after. These two steps, when executed after binary segmentation and after the refitting method described from Section 2.4.2, ensure that the first stage estimates are more robust to overestimating $J$.

To illustrate the rate of convergence, intelligent sampling was applied to a data sequence of length $N$ varying from $10^5$ to $10^{7.5}$, evenly on the log scale, with the number of change points being $J \approx \log_{10}(N)^2$. The change point location and the signal levels were randomly generated:

- The spacings $(\tau_1, \tau_2 - \tau_1, \tau_3 - \tau_2, ...., N - \tau_J)$ were generated as the sum of the constant $\frac{N}{1.5J}$ and the Multinom $\left(N - \frac{(J+1)N}{1.5J}, (p_0, ..., p_J)\right)$ distribution
  - $(p_0, ..., p_J)$ is generated as the consecutive differences of the order statistics of $J + 1$ Unif(0,1) random variables
- The signals were generated as a Markov chain with $\nu_0$ initialized as 0, and iteratively, given $\nu_i$, $\nu_{i+1}$ is generated from a density proportional to $f(x) = \exp(-0.3(x - \nu_i - \underline{\Delta}))1(\nu_i + \underline{\Delta} \leq x \leq M) + \exp(0.3(x + \nu_i - \underline{\Delta}))1(\nu_i - \underline{\Delta} \geq x \geq -M)$, where $M$ was taken to be 10 and $\underline{\Delta}$ was taken to be 1.

For each of 10 values of $N$, 50 configurations of change points and signals were generated, and on each of those configurations 40 datasets with iid $N(0,1)$ error terms simulated, and intelligent sampling was performed on each simulated dataset with binary segmentation taken at stage 1 on a subsample of size roughly $N_1 = 50\sqrt{N}$ and thresholding parameter $\zeta_N = N_1^{0.2}$, a valid choice according to Theorem II.6. Additionally the drop steps (D1) and (D2) were applied right after binary segmentation (with $\delta_D = 15$ and $\Delta_D = 0.5$), the re-estimation procedures (ISM4-1) to (ISM4-3) were run, and the drop steps (D1) and (D2) were applied again. With this setup, the stage 1 binary segmentation was accurate in determining the correct value of $J$ over 99% of times for all $N$. In the second stage we let the width of the sampling interval around $\hat{\tau}_j^{(1)}$ be the $Q_{\tilde{\Delta}_j}\left(1 - \frac{0.01}{\hat{J}}\right)$ where $\tilde{\Delta}_j = \frac{\hat{\nu}_j^{(1)} - \hat{\nu}_{j-1}^{(1)}}{\hat{\sigma}}$, for $j = 1, \ldots, \hat{J}$. After the second stage of intelligent sampling, the maximum deviations $\max_{j=1,\ldots,J} |\lambda_2(\tau_j, \hat{\tau}_j^{(2)})|$ were recorded. The running time of intelligent sampling was also recorded, and we compared it to the running time of binary segmentation on the full data (only the BinSeg procedure itself, without steps (D1) and (D2)). For the latter, we ran 100 iterations for each $N$, 2 runs per configuration of parameters. The results are depicted in Figure 2.9 and are in accordance with the theoretical investigations: the quantiles of $\max |\hat{\tau}_j^{(2)} - \tau_j|$ scale sub-linearly with $\log(J)$ (which is $\sim \log(\log(N))$ in this setup where $J \sim \log^2(N)$) as predicted in Theorem II.4, and the computational time of intelligent sampling scales in the order of $\sqrt{N}$ compared to the order $N$ computational time of using BinSeg on the entire dataset.

A second set of simulation experiments was used to illustrate the asymptotic distribution of the change point deviations. We considered four scenarios, each with $N = 10^7$ and 55 change points, which was the maximum number of change points used in the last simulation setting.

(Setup 1): one set of signal and change point locations generated as in the previous set of simulations, with i.i.d. $N(0,1)$ error terms

(Setup 2): change points evenly spaced apart with signals 0,1,0,1,.., repeating, and i.i.d. $N(0,1)$ error terms

(Setup 3): change points evenly spaced part with 0,1,0,1,... repeating signals, and error terms generated as $\varepsilon_i = \frac{\varepsilon_i^* + 0.5\varepsilon_{i+1}^* + 0.25\varepsilon_{i+2}^*}{\sqrt{1^2+0.5^2+0.25^2}}$ for all $i = 1, ..., N$, where the $\varepsilon_i^*$'s are generated as i.i.d. $N(0,1)$;

(Setup 4): change points evenly spaced with signals 0,1,0,1,..., and error terms generated from an AR(1) series with parameter 0.2, and each marginally $N(0,1)$.

For all 4 cases the first stage of intelligent sampling was performed identically as for the previous set of simulations, and with the same tuning parameters. At the second stage, first stage subsample points were omitted for data with setups 1 and 2, but not for setups 3 and 4. At stage 2, the subsampling intervals had widths that equal the $1 - \frac{0.01}{\hat{J}}$ quantile of the $L$-type distributions. From 2500 iterations on each of the 4 simulation setups, the distributions of the maximum deviations (maximum of $|\lambda_2(\tau_j, \hat{\tau}_j^{(2)})|$ for the first two setups and $|\hat{\tau}_j^{(2)} - \tau_j|$ for the other two setups) are seen to match well with their predicted asymptotic distributions. To illustrate the convergence of the individual change point estimates, we also show that the distribution of the 27th change point matches with the $L$-type distributions seen in Proposition 3.

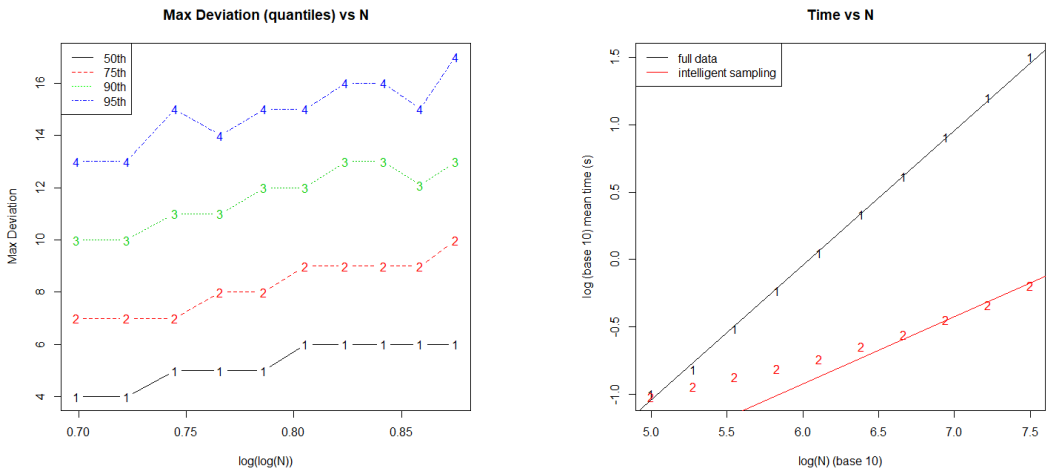


Figure 2.9: **Left:** Quantiles of the max deviations versus $\log(\log(N))$, which is the same order as $\log(J)$. Over the observed regime of parameters, the maximal deviation scales with $J$, as was predicted by Theorem II.4. **Right:** Log-log plot of mean computational time when using intelligent sampling to obtain the final change-point estimates at stage two, and using BinSeg on the full data to construct change-point estimates, with reference lines of slope 1 (black) and 0.5 (red) respectively. To give some sense of the actual values, for $N = 10^{7.5}$ the average time for intelligent sampling vs full data were, respectively, 0.644 and 31.805 seconds.

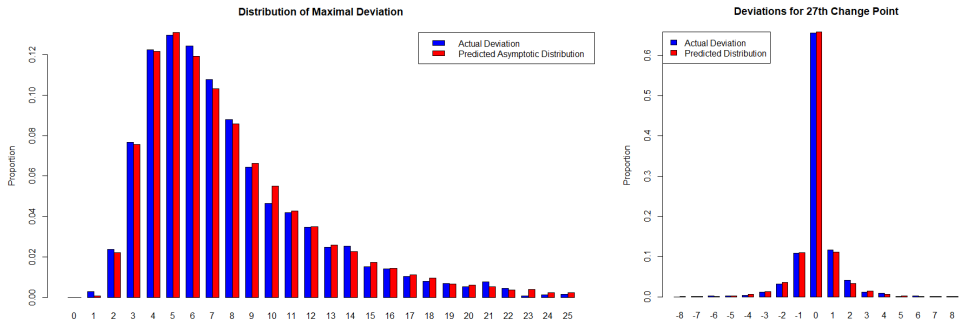


Figure 2.10:
Distributions of $\max_{1\leq j\leq 55}\lambda_2\left(\tau_j,\hat{\tau}_j^{(2)}\right)$ (**Left**) and $\lambda_2\left(\tau_{27},\hat{\tau}_{27}^{(2)}\right)$ (**Right**) from simulations of setup 1.

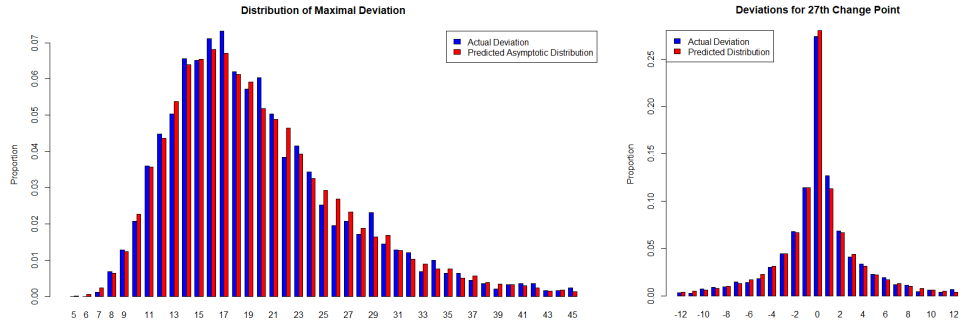


Figure 2.11:
Distributions of $\max_{1\leq j\leq 55}\lambda_2\left(\tau_j,\hat{\tau}_j^{(2)}\right)$ (**Left**) and $\lambda_2\left(\tau_{27},\hat{\tau}_{27}^{(2)}\right)$ (**Right**) from setup 2.

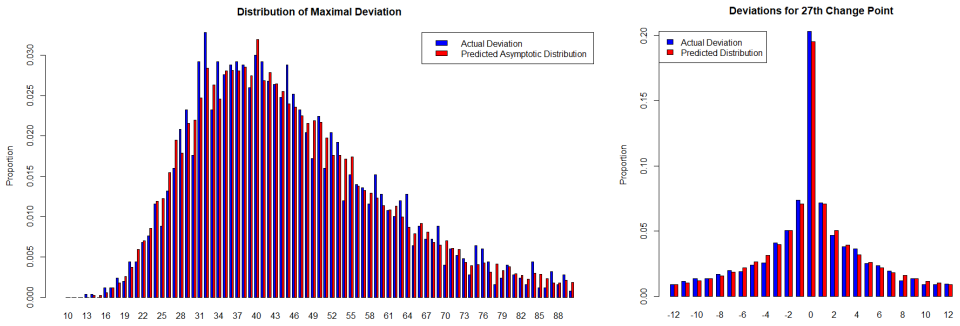


Figure 2.12: Distributions of $\max_{1\leq j\leq 55}\left|\tau_j-\hat{\tau}_j^{(2)}\right|$ (**Left**) and $\left|\tau_{27}-\hat{\tau}_{27}^{(2)}\right|$ (**Right**) from setup 3

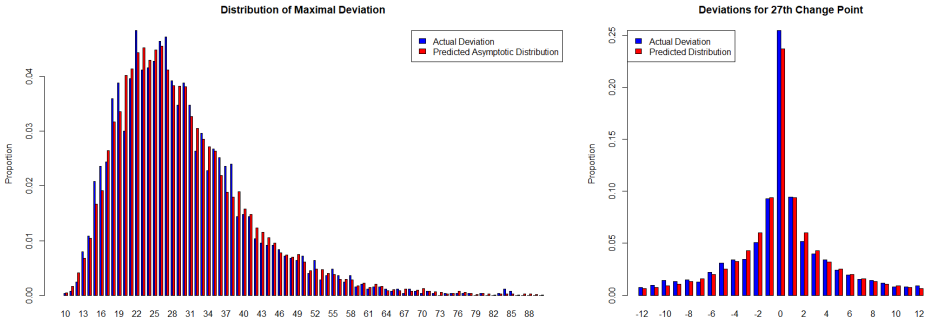


Figure 2.13: Distributions of $\max_{1\leq j\leq 55}\left|\tau_j-\hat{\tau}_j^{(2)}\right|$ (**Left**) and $\left|\tau_{27}-\hat{\tau}_{27}^{(2)}\right|$ (**Right**) from setup 4.

The distribution of the deviations for setup 1 is the least spread out, with the primary reason that the jump between signals is randomly generated but lower bounded by 1 while the other 3 setups have signal jumps all fixed at 1. Setups 3 and 4 have the most spread out distributions, as the dependence among the error terms causes the estimation to be less accurate but only up to a constant and not an order of magnitude. Nevertheless, in all 4 setups, the change point estimates behave very closely to what was predicted in Theorems II.3 and II.5 (for the first two scenarios)

and Proposition 3 (for the third and fourth scenarios).

**Implementation in a heteroscedastic error model:** We further explored the validity of Proposition 3 by looking at a case with heteroscedatic errors. We again generated a data sequence of length $N = 10^7$ with 55 evenly spaced changed points and signals of $0, 1, 0, 1, 0, 1, \ldots, 1$. Instead of generating error terms from a stationary series, we, instead, generated them as independent segments of Gaussian processes as follows. For $j = 1, 2, 3, \ldots,$

1. from $\tau_{4j}$ to $\frac{\tau_{4j+2}+\tau_{4j+1}}{2}$ the errors are iid $N(0, 1)$;
2. from $\frac{\tau_{4j+2}+\tau_{4j+1}}{2}$ to $\frac{\tau_{4j+3}+\tau_{4j+2}}{2}$ the errors are $\varepsilon_i = \frac{\varepsilon_i^* + 0.5\varepsilon_{i+1}^* + 0.25\varepsilon_{i+2}^*}{\sqrt{1^2+0.5^2+0.25^2}}$ where the $\varepsilon_i^*$'s are iid $N(0, 1)$ (and will be treated as a generic iid $N(0, 1)$ sequence from here on;)
3. from $\frac{\tau_{4j+3}+\tau_{4j+2}}{2}$ to $\tau_{4j+3}$, error terms are $\varepsilon_i = 0.5 \cdot \frac{\varepsilon_i^* + \varepsilon_{i+1}^* + \varepsilon_{i+2}^* + \varepsilon_{i+3}^*}{\sqrt{4}}$;
4. from $\tau_{4j+3}$ to $\tau_{4j+4}$ the error terms are $\varepsilon_i = 0.7 \cdot \frac{\varepsilon_i^* + \varepsilon_{i+3}^*}{\sqrt{2}}$;

and the error terms generated in each stretch are independent of those in any other stretch. This creates a situation where around $\tau_{4j+1}$ the error terms are iid $N(0, 1)$, around $\tau_{4j+2}$ the error terms are stationary, and around $\tau_{4j+3}$ and $\tau_{4j+4}$ the error terms are stationary to the left and to the right, but their autocorrelation and marginal variances change at the change points. With the same intelligent sampling procedure as setups 3 and 4, and the same tuning parameters, we ran 2000 replicates of this setup and recorded the $\hat{\tau}_j^{(2)} - \tau_j$ values for $j = 1, \ldots, 55$.

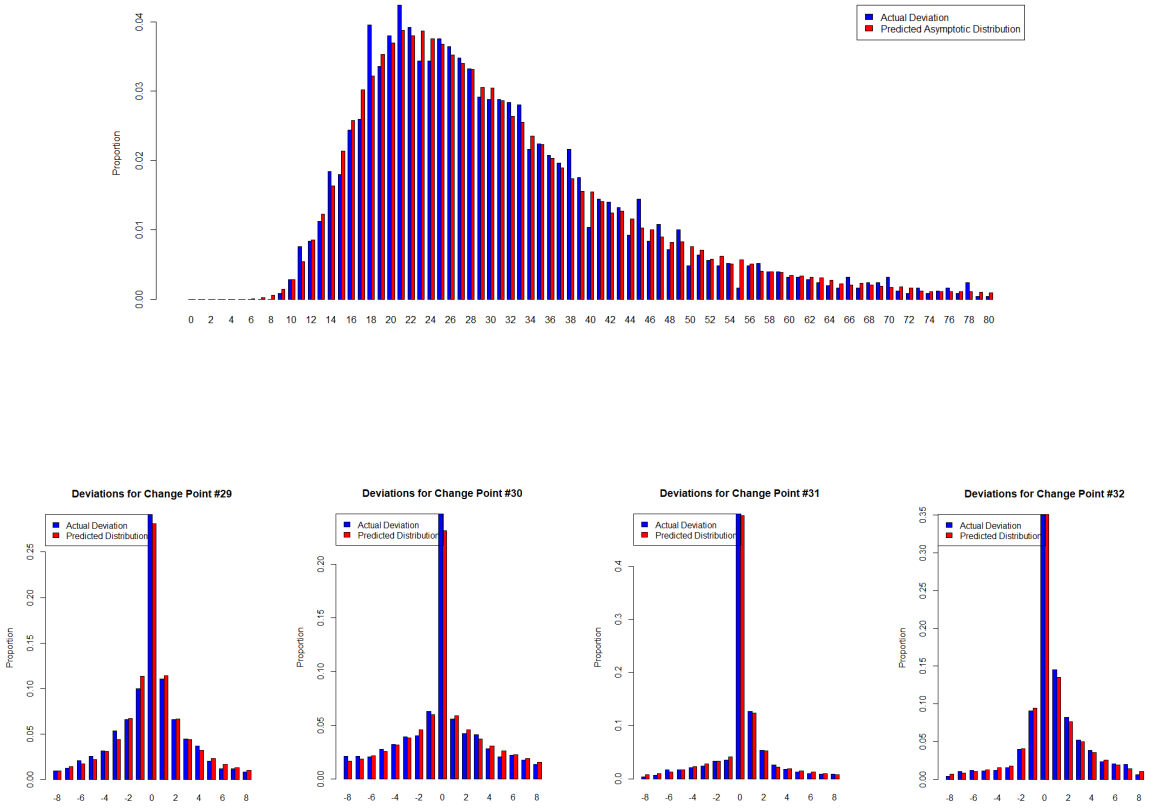


Figure 2.14: **Top:** Predicted and actual distribution of the maximal deviation $\max_{1\leq j\leq 55}|\hat{\tau}_j^{(2)} - \tau_j|$. **Bottom, Left to Right:** Predicted and actual distributions of the individual deviations $|\hat{\tau}_j^{(2)} - \tau_j|$ for $j = 29, 30, 31,$ and 32.

Results from the simulation are very consistent with Proposition 3. Even for change points which have different distributions of error terms to the left and right, the deviations match up very closely with the stated asymptotic distributions.

## 2.8 Real Data Application

The effectiveness of the proposed intelligent sampling procedure is illustrated on an Internet traffic data set, obtained from the public CAIDA repository

[3] that contains traffic traces from an OC48 (2.5 Gbits/sec) capacity link. The trace under consideration contains traffic for a two hour period from large west coast Internet service provider back in 2002. The original trace contains all packets that went through the link in an approximately 2 hour interval, but after some aggregation into bins of lengh 300 microseconds, the resulting data sequence comprises of $N = 1.5 \times 10^7$ observations. After applying a square-root transformation, a snapshot from this sequence is depicted in Figure 2.15 and some of its statistical characteristics in Figure 2.16, respectively.

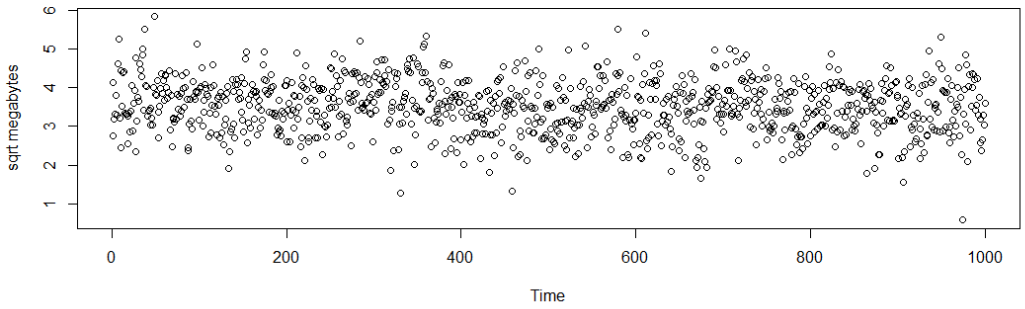


Figure 2.15: First 5000 time points of the data after a square root transformation.

[3]http://data.caida.org/datasets/passive/passive-oc48/20020814-160000.UTC/pcap/

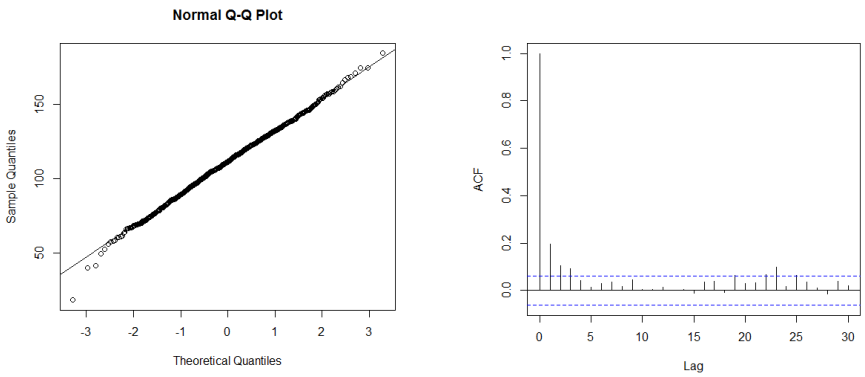


Figure 2.16: QQ plot and estimated ACF of first 5000 points of data set, after square root transformation, note the normality of the data after a square root transformation.

It can be seen that the data are close to marginally normally distributed, while their autocorrelation decays rapidly and essentially disappears after a lag of 10. Similar exploratory analyses performed for multiple stretches of the data leads to similar conclusions. Hence, for the remainder of the analysis, we work with the square-root transformed data and model them as a short range dependent sequence.

To illustrate the methodology, we used an *emulation* setting, where we injected various mean shifts to the mean level of the data of random durations, as described next. This allows us to test the proposed intelligent sampling procedure, while at the same time retaining all features in the original data.

In our emulation experiments, we posit that there are two types of disruptions, short term spikes that may be the result of specific events (release of a software upgrade, a new product or a highly anticipated broadcast) and longer duration disruptions that may be the result of malicious activity *Carl et al.* (2006); *Kallitsis et al.* (2016). To emulate these scenarios, at random intervals $[V_i, W_i]$ we increased the signal in the data sequence by a randomly generated constant $\Delta_i$, changing the data as $Y_j \leftarrow Y_j + \Delta_i \cdot 1_{[V_i,W_i]}(j)$, as follows:

(sig-1): A set of 31 stretches $(V_{1,i}, W_{1,i})$ for $i = 1, \ldots, 31$ were created by first generating $(W_{1,1}, W_{1,2} - W_{1,1}, W_{1,3} - W_{1,2}, \ldots, W_{1,30} - W_{1,29}, N - W_{1,30})$ from the $1.5 \times 10^5$+Multinom$(N - 31 \times 1.5 \times 10^5, (p_0, \ldots, p_{30}))$ distribution, conditional on $(p_0, \ldots, p_{120})$ which is generated as the order statistics of Unif$(0, 1)$. Then each $V_{1,i}$ are taken by generating each $W_{1,i} - V_{1,i}$ as 75000+Binom$(W_{1,i} - W_{1,i-1} - 75000, 0.5)$, and the increase in signals $\Delta_{1,i}$'s are generated from a Unif$(1.3\hat{\sigma}, 2\hat{\sigma})$ where $\hat{\sigma}$ is the standard deviation of the data.

(sig-2): Stretches $(V_{2,i}, W_{2,i})$ for $i = 1, \ldots, 201$ were independently generated by setting $(W_{2,1}, W_{2,2} - W_{2,1}, \ldots, N - W_{2,200})$ from the 50050+Multinom$(N - 201 \times 50050, (p_0, \ldots, p_{201}))$ conditioned on $(p_0, \ldots, p_{120})$ which is generated as the order statistics of Unif$(0, 1)$. Each $W_{2,i} - V_{2,i}$ is generated as 50+Binom$(W_{2,i} - W_{2,i-1} - 50000, 0.0001)$, and the increases in signals as $\Delta_{2,i} \sim$Unif$(10\hat{\sigma}, 15\hat{\sigma})$

This scheme randomly places a fixed number of stretches of traffic increases (a combined value of 232 in fact), without placing the stretches too close together. Stretches from (sig-1) emulate longer, milder increases of a bump in the data sequence, as each $W_{1,i} - V_{1,i} \geq 75,000$, while stretches from (sig-2) emulate short but more dramatic increases, as each $W_i - V_i$ is guaranteed to be higher than 50 but not likely to be much higher. Both types of traffic increases can occur when looking for increase in user traffic or attacks by third parties. A depiction of a segment of the data with the emulated signal is given in Figure 2.17.

As mentioned in the introduction, the main objective of the proposed methodology is to identify *long duration, persistent shifts* in the data sequence using a limited number data points; in the emulation scenario used, this corresponds to change points induced by sig-1, while we remain indifferent to those induced by sig-2.[4]

The two-stage intelligent sampling procedure was implemented as follows: (i) the

[4]We note that the theoretical development does not include spiky signals. Nonetheless, we included spiky signals in our emulation to mimic the pattern of internet traffic data. As will be seen later, our method is quite robust to the presence of this added feature.

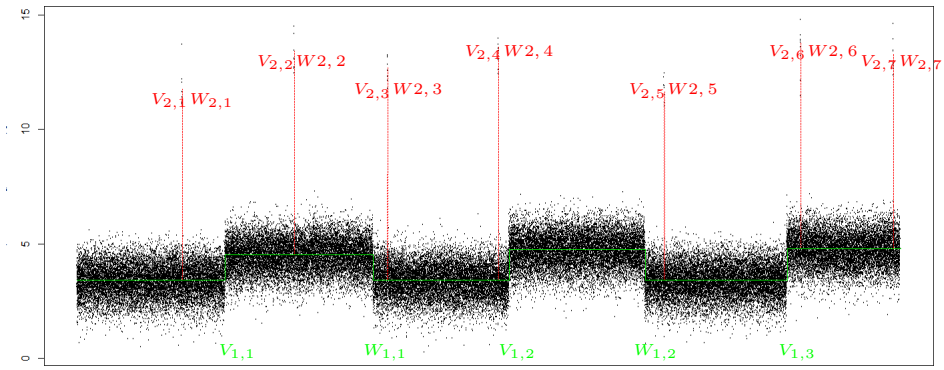


Figure 2.17: Example of emulated data. The intervals $[V_{1,i}, W_{1,i}]$ emulate persistent stretches of mild increase in traffic, while the intervals $[V_{2,i}, W_{2,i}]$ emulate very short stretches of high traffic increase.

first stage subsample comprised observations 100 time points apart in the original data sequence. BinSeg with thresholding parameter $\zeta_{N_1} = N_1^{0.2} = 150,000^{0.2}$ was employed, followed by steps (D1) and (D2) from Section 2.7 with $\delta_D = 15$ and $\Delta_D = 0.5$, and calibration from Section 2.4.2 applied with a different subsample, and again an application of steps (D1) and (D2). (ii) For each $j$, the second stage interval surrounding $\hat{\tau}_j^{(1)}$ was chosen to have half width $Q_{\tilde{\Delta}_j}\left(1 - \frac{0.01}{\hat{J}}\right)$ where $\tilde{\Delta}_j = \frac{\hat{\nu}_j^{(1)} - \hat{\nu}_{j-1}^{(1)}}{\hat{\sigma}}$. A stump model was then fitted to the data in each second stage interval to obtain the final estimates of the change-points and the final (2nd stage) CIs were constructed.

To assess the accuracy, we calculated the coverage proportion of the 90%, 95%, and 98% level confidence intervals over different emulation settings. To construct these confidence intervals we had to randomly generate data sequences with identical distribution structure as the data (which would give us a random sample of $L$-type distributions and their quantiles). We generated these sequences as marginally normal random variables, with marginal standard deviation the same as the sample sd of the first 50,000 points of the $1.5 \times 10^7$ length data. Finally, the ACF of the generated

series was matched with the sample ACF of the first 50,000 points up to a lag of 20: we first generated vectors of iid normal variables, then multiplied them with the Cholesky square root of the band matrix created with the sample ACF (bandwidth of this matrix is 20, and non-zero entries taken from the first 20 values of the sample ACF).

Intelligent sampling exhibits satisfactory performance: among all 61 change points corresponding to sig-1, the lowest coverage probability for the 90%, 95%, and 98% nominal confidence intervals were 0.892, 0.926, and 0.954 respectively, while average coverage probabilities were around 0.906, 0.945, and 0.969, respectively. On the other hand, for change points induced by sig-2, the average coverage probability was lower than 0.002 even for the 98th confidence interval. However, since the focus of intelligent sampling is on long duration persistent signals, missing the spiky signals is of no great consequence. In terms of computational burden, the average emulation setting utilized 3.46% of the full dataset, requiring an average time of 1.43 seconds to perform the estimation. We note that Table 2.5 corresponds to the length of this dataset in Section 2.8. As we are concerned with SNR between 1.3 and 2, the percentage of data used in our emulation experiment is quite consistent with the numbers presented in that table.

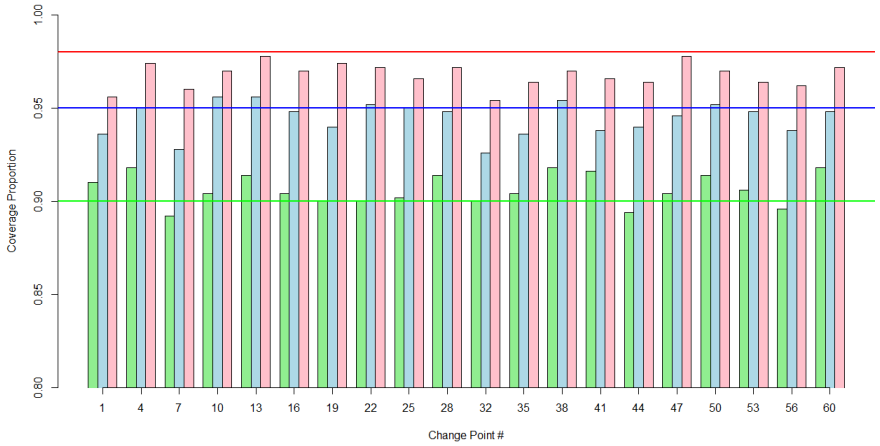


Figure 2.18: Coverage proportions, the proportion of time when the change point was covered by some confidence interval, for the 90% level (green bars), 95% level (blue bars), and 98% level (red bars) within the 500 iterations, for a select number of 20 change points (change point # 2 is always the second one in order, # 3 is the third in order, etc). Horizontal reference lines are at 0.9 (green), 0.95 (blue), and 0.98 (red).

## 2.9 Concluding Remarks and Discussion

This chapter introduced sampling methodology that reduces significantly the computational requirements in multi-change point problems, while not compromising on the statistical accuracy of the resulting estimates. It leverages the *locality principle*, which is obviously at work in the context of the classical signal-plus-noise model employed throughout this study. A natural extension, further enhancing the scope of the methodology, would deal with a piece-wise linear signal plus noise model with discontinuities between the linear stretches. We expect that the locality principle naturally extends to such a setting, based on prior work together with numerical results in a related problem *Lan et al.* (2009) in a design setting provides. Extensions to

problems involving multiple (potentially high-dimensional) data sequence produced by cyber-physical systems equipped with a multitude of sensors monitoring physical or man-made phenomena are of obvious interest.

The focus in this chapter has primarily been on a two-stage procedure, which is easiest to implement in practice and suitable for many applications. Nevertheless, as illustrated in Section 2.5, in specific settings involving data sets of length exceeding $10^{10}$ points, a multi-stage procedure may be advantageous,

A key technical requirement for intelligent sampling is that the procedure used to obtain the 1st stage estimates needs to exhibit consistency properties, e.g. (2.14). The choice of binary segmentation in our exposition, or its wild binary segmentation variant (which modifies BinSeg by computing the cusum statistics on an additional number of random intervals) presented in detail in Section A.4 of the supplementary material, is due to their computational attractiveness and the fact that they readily provide consistent estimates of the number of change points and their locations. Nevertheless, there are other methods that fit the bill, as discussed next.

Two popular methods used for models defined as in (2.11) are the estimation of multiple structural breakpoints introduced in *Bai and Perron* (1998) and PELT as described in *Killick et al.* (2012). The method described in *Bai and Perron* (1998) does give consistent estimates, but only under the much stricter condition that $J$ is a constant and there exists values $\beta_1, \ldots, \beta_J \in (0, 1)$ such that $\tau_j = [\beta_j N]$ for all $j = 1, \ldots, J$ and $N$. Further, to run the actual procedure would require the use of dynamic programming which is computationally expensive ($O(N^2)$ time). With the PELT procedure, the implementation itself runs in a more manageable $O(N)$ time; however, this works under the very different Bayesian setting where the spacings $\tau_{j+1} - \tau_j$ are iid generated from some distribution. Further, PELT was built upon a procedure described in *Yao* (1984), which examines another Bayesian model where every point $\{1, \ldots, N\}$ has a probability $p$ of being a change point, and the development did not

go into details regarding rates of convergence of the change point estimates. Due to the theoretical and computational restrictions of the multiple structural breakpoints method and the differing framework under which PELT works, we focused our analysis on binary segmentation.

We also mention the SMUCE procedure, introduced in *Frick et al.* (2014), where lower probability bounds for the events $\mathbb{P}[\hat{J} \neq J]$ and $\mathbb{P}[\max_{j=1,\ldots,J} \min_{i=1,\ldots,\hat{J}} |\hat{\tau}_i - \tau_j| \leq c_N]$, for any sequence $c_N$, were derived. These results can be combined to yield $\mathbb{P}[\hat{J} = J; \quad \max_{j=1,\ldots,J} |\hat{\tau}_j - \tau_j| \leq c_N] \to 1$ under certain restrictions and for some sequences $c_N$ that are $o(\delta_N)$, and therefore could be used in the first stage of intelligent sampling. SMUCE has the flexibility of working for a broader class of error terms [5] but as was stated in *Frick et al.* (2014), the procedure involves dynamic programming which runs in $O(N^2)$ time. This last point is less of an issue for a modified version of SMUCE designed for iid Gaussian errors with heterogeneous variances. H-SMUCE, in *Pein et al.* (2016), could run the procedure in as low as $O(N)$ time in some cases. Overall, SMUCE could be used as the first part of intelligent sampling, and the regimes of $\delta_N$ and restrictions on the subsample size $N_1$ needed for intelligent sampling to be consistent could be fleshed out in a similar manner as in this paper. However, as BinSeg and WBinSeg are somewhat easier to implement computationally, we chose to perform our analysis with them instead.

While we work with i.i.d. Gaussian error terms in this paper, our simulation results indicate that for non-Gaussian i.i.d. errors or dependent error terms (with or without Gaussian distributions), the deviation of our estimator behaves like the $L$-type distributions. This suggests that our results could be extended to broader classes of error terms, and future work can consider models incorporating errors with dependence structures and/or with non-Gaussian distributions. Extending Theorem II.3 to these settings would require an in-depth investigation into probability bounds

[5] Some results apply when the errors are iid from a general exponential family.

on the argmin of a drifted random walk with non-i.i.d and/or non-Gaussian random components dealt with in Section A.3 of the Supplement. We speculate that this work would be more amenable to rigorous analysis when the tail probabilities of the error terms decay exponentially (similar to Gaussian distributions) and when the dependence is local , e.g. $m$-dependence, where each error terms is only correlated with its $m$ neighbors to the left and the right.

In conclusion, any procedure used at stage 1 of intelligent sampling puts restrictions on the model specifications, as consistent second stage estimators cannot be obtained if the first stage procedure is not consistent. Established results for BinSeg, as in *Venkatraman* (1992) and *Fryzlewicz et al.* (2014), consider only the i.i.d Gaussian framework. Extending the BinSeg based approach to a more flexible class of error terms therefore requires theoretical exploration of BinSeg's properties beyond Gaussian errors, or using alternative methods at stage 1 which do not need the Gaussian error framework, e.g. *Bai and Perron* (1998) and *Frick et al.* (2014).

# CHAPTER III

# Change Planes in Growing Dimensions

## 3.1 Introduction

The simple regression model typically assumes a single uniform relationship between the covariate and the response, in the form of a single $\beta$ parameter that relates $X$ and $Y$. In practice, there does not have to be a single underlying relationship between every covariate and response pair; for instance, there could be several sub-populations, each with a different $\beta_j$ parameter. To account for heterogeneity among the covariate response-relationship, some common techniques include mixed linear models and fitting different models among each sub-population, which correspond to a supervised classification setting where the true groups are known.

A different challenge is faced when the true subgroups among covariates are not known, and thus regression must occur after or concurrently with classification. To study this problem, it is typical to consider the case where the covariates fall into two unknown subgroups which can be delineated, or notation-wise, $E[Y_i|X_i]$ could be one of two linear functions depending on whether $X_i$ falls within subgroup 1 or 2. Earlier treatments of this problem, such as *Hansen* (2000), study a linear thresholding model where membership between the two subgroups is determined by whether a real-valued observed variable $Q$ falls within either side of an unknown parameter

$\gamma$. More recently there has been focus on an SVM-type model where the membership is determined by which side $X$ falls with respect to an unknown hyperplane $\theta_0$; for a more explicit example, *Wei and Kosorok* (2014) extended the linear thresholding model found in *KANG et al.* (2007) to propose the model

$$Y = \mu_1 \cdot 1_{X^T\theta_0 \geq 0} + \mu_2 \cdot 1_{X^T\theta_0 < 0} + \varepsilon. \tag{3.1}$$

This model and others with a similar structure, called the change plane model, has seen a variety of applications, utilized to model treatment effect heterogeneity in drug treatment (*Imai et al.* (2013)), model sociological data on voting and employment (*Imai et al.* (2013)), and to model cross country growth regressions in econometrics (*Seo and Linton* (2007)).

In terms of academic study this model has also seen a number of extensions and developments. Some of these studies have paralleled the development of the one dimensional change point problem. Namely, *Fan et al.* (2017) studied the change plane in a statistical test setting, with the null being no hyperplane thresholding and the alternative being that there is, by proposing a test statistic, studying its asymptotic distribution, and giving sample size recommendations. At around a similar time, *Li et al.* (2018) extended the single hyperplane thresholding problem to a multiple thresholding problem by considering a model with parallel change planes. To extend the change plane model in such ways is intuitive, as one can see it as discontinuity estimation in higher dimensions. However, not all work has continued in this vein.

As a non-standard thresholding problem in multiple dimensions, the change plane problem is a mathematical curiosity on top of its practical applications. At the same time, it is at its heart a classification and regression problem. Both of these subjects

have seen a plethora of modern work in the growing/high dimensional framework (see *Tibshirani* (1996) and *Fan and Fan* (2008) for examples). Even in the specific case of regression with multiple subpopulations, the high dimensional direction has been recently explored in the supervised case (see *Dondelinger et al.* (2016) for details). Meanwhile, the growing and high dimensional scenario for the change plane model is still unexplored, despite motivations including the aforementioned extensions in similar fields, practical relevance, and the theoretical attractiveness of being a non-standard problem in high dimensions.

Finally, a third avenue of work has been focused on the methodological implementation. As a model with discontinuous features, the optimization problems involved are usually not differentiable and therefore cannot be directly dealt with using standard methods. To estimate the multi-dimensional change plane parameter would require different strategies. Consequently, both *Li et al.* (2018) and *Seo and Linton* (2007) have proposed optimizing smoothed proxy expressions in place of the discontinuous true expressions, as well as deriving rates for these proxy estimators.

In this chapter we will delve into the analysis of the change plane problem for growing dimensions. We will first focus on a canonical problem very similar to (3.1), with one main difference being the response is distributed as a Bernoulli random variable conditional on the covariate. As change plane in growing dimensions is an unexplored area, our main focus will to establish basic asymptotic results. First, consistency and rates of convergence will be established for this model on a low but growing dimensional scenario. Later, results will be covered for the high dimensional case, with the estimator chosen after an $\ell_0$ penalization. Afterwards, extensions in the continuous response model will also be covered.

## 3.2 Binary Response Model

A basic change plane model would involve a covariate and response ($X$ and $Y$ respectively) where the mean of the response, differs on either side of an unknown hyperplane. In mathematical notation, consider the case where

$$\mathbb{E}\left[Y|X\right] = \alpha_0 \cdot 1(X^T\theta_0 \leq 0) + \beta_0 \cdot 1(X^T\theta_0 > 0) \tag{3.2}$$

for some parameters $\alpha_0 \in \mathbb{R}$, $\beta_0 \in S^{d-1}$ (with $S^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ being the unit sphere in $d$ dimensions), and $\theta_0 \in \mathbb{R}^d$, the latter being of primary interest for estimation. Due to our interest in analysis for growing dimensions, the dimension of the covariates $d = d_n$ can increase without bound. As a starting model, we will consider the case where $Y$ is a binary response variable taking on values of either 0 or 1 (i.e. $Y$ is Bernoulli with parameter $\alpha_0$ or $\beta_0$ conditional on $X$), and hence we place the restriction $0 < \alpha_0 < \beta_0 < 1$.

## 3.3 Least Squares Estimator

As a starting point for analysis on the asymptotic behavior of the problem, we first consider estimating for $(\alpha_0, \beta_0, \theta_0)$ using the often well-behaved least squares estimator:

$$(\hat{\alpha}^{(sq)}, \hat{\beta}^{(sq)}, \hat{\theta}^{(sq)}) := \underset{\substack{0\leq\alpha<\beta\leq 1 \\ \theta\in S^{d-1}}}{\arg\min} \mathbb{P}_n \left[\left(Y - \alpha \cdot 1(X^T\theta \leq 0) - \beta \cdot 1(X^T\theta > 0)\right)^2\right]. \tag{3.3}$$

As well shall see, these estimators are not only consistent, but also converge at a min-max rate modulo a log factor. To derive these results, we first consider the following assumptions

**Assumption A1**: There exist positive constants $a^-$ and $a^+$ such that for every

$n$,

$$a^- \|\theta_1 - \theta_2\|_2 \leq P\left(X^T\theta_1 < 0 < X^T\theta_2\right) \leq a^+ \|\theta_1 - \theta_2\|_2 \tag{3.4}$$

for every $\theta_1, \theta_2 \in S^{d-1}$.

**Remark 15.** *Note that Assumption A1 implies that for every* $n$, $P\left(X^T\theta_1 > 0, X^T\theta_2 > 0\right) > 0$ *whenever* $\theta_1 \neq \theta_2$. *This can be seen by plugging* $(-\theta_2)$ *for* $\theta_2$ *in (3.4).*

**Assumption A2**: $d = d_n = o(n)$

Assumptions A1 pertains to the distribution of covariate $X$, and without it one may construct counterexamples where the estimates exhibit undesirable behavior. For instance, consider the case where there exists open sets $U_n$ such that $\{x \in \mathbb{R}^d : x^T\theta_0 = 0\} \subseteq U_n$ and $\mathbb{P}(X \in U_n) = 0$, then the least squares estimates may not even be consistent. On the other hand, when it is satisfied, we can mathematically derive a lower bound for the excess loss

$$\begin{aligned} &P\left[\left(Y - \alpha \cdot 1(X^T\theta \leq 0) - \beta \cdot 1(X^T\theta > 0)\right)^2\right] - \\ &P\left[\left(Y - \alpha_0 \cdot 1(X^T\theta_0 \leq 0) - \beta_0 \cdot 1(X^T\theta_0 > 0)\right)^2\right]. \end{aligned} \tag{3.5}$$

Next, Assumption A2 lets us work in a low dimensional framework. Mathematically, this will allow the use of a VC dimension argument to obtain a Glivenko-Cantelli type result. Combining these two results implies the consistency of the least squares estimators.

**Theorem III.1.** *If Assumptions A1 and A2 are satisfied, then* $\hat{\alpha}^{(sq)} - \alpha_0$, $\hat{\beta}^{(sq)} - \beta_0$, *and* $\|\hat{\theta}^{(sq)} - \theta_0\|_2$ *all converge to 0 in probability.*

*Proof.* See Section B.0.0.1. □

Next, results from M-estimation can be used to derive rates of convergence. The ingredients needed to apply this result are a lower bound on the difference in (3.5), and a bound on a modulus of continuity. The former can be provided, again, due to Assumption A1, while the latter can be derived as a result of working with a function class with a finite envelope function, and a VC dimension bounded by a scalar times $d$.

**Theorem III.2.** *Assuming conditions A1 and A2,* $(\hat{\alpha}^{(sq)}-\alpha_0)^2$, $(\hat{\beta}^{(sq)}-\beta_0)^2$, $\|\hat{\theta}^{(sq)}-\theta_0\|_2$ *are* $O_p\left(\frac{d}{n}\log\left(\frac{n}{d}\right)\right)$.

*Proof.* See Section B.0.0.3. □

**Remark 16.** *We note that in the case where* $\frac{d}{n}\log\left(\frac{n}{d}\right)=o(n)$, *estimators for* $(\alpha_0,\beta_0)$ *with faster optimal rate can be easily obtained. Take the following as an estimator for* $\alpha_0$*:*

$$\hat{\alpha}=\frac{\frac{1}{n}\sum_{i=1}^n \mathbb{1}\left(Y_i=1, X_i^T\hat{\theta}^{(sq)}\le 0\right)}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}\left(X_i^T\hat{\theta}^{(sq)}\le 0\right)}. \tag{3.6}$$

*The denominator is*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{1}\left(X_i^T\hat{\theta}^{(sq)}\le 0\right)=\frac{1}{n}\sum_{i=1}^n \mathbb{1}\left(X_i^T\theta_0\le 0\right)+\frac{1}{n}\sum_{i=1}^n \mathbb{1}\left(X_i^T\hat{\theta}^{(sq)}\le 0, X_i^T\theta_0>0\right) \tag{3.7}$$

*with the first term equal to* $P(X^T\theta_0\le 0)+O_p(1/\sqrt{n})$, *as the first term is a Binomial variable with mean* $P(X^T\theta_0\le 0)$ *and variance at most* $\frac{1}{4n}$, *and the second term is* $O_p\left(\left(\frac{d}{n}\log\left(\frac{d}{n}\right)\right)\vee\frac{1}{\sqrt{n}}\right)$. *To see the latter statement, the second term, conditional on* $\hat{\theta}^{(sq)}$, *is Binomial with mean* $\mathbb{P}\left(X^T\hat{\theta}^{(sq)}\le 0<X^T\theta_0\right)\le a^+\|\theta_0-\hat{\theta}^{(sq)}\|_2$ *and variance no greater than* $1/n$. *Hence, using Chebychev's inequality, for any fixed* $\epsilon>0$ *we have*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \mathbb{1}\left(X_i^T\hat{\theta}^{(sq)}\le 0, X_i^T\theta_0>0\right)\le a^+\|\hat{\theta}^{(sq)}-\theta_0\|_2+\sqrt{\frac{2}{\epsilon}}\sqrt{\frac{1}{n}}\middle|\hat{\theta}^{(sq)}\right]\ge 1-\frac{\epsilon}{2},$$

*and taking expectations leads to*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(X_i^T\hat{\theta}^{(sq)}\leq 0, X_i^T\theta_0>0\right)\leq a^+\|\hat{\theta}^{(sq)}-\theta_0\|_2+\sqrt{\frac{2}{\epsilon}}\sqrt{\frac{1}{n}}\right]\geq 1-\frac{\epsilon}{2}.$$

*Combine this with the fact that there exists a constant $C>0$ such that*

$$\mathbb{P}\left[a^+\|\hat{\theta}^{(sq)}-\theta_0\|_2\leq C\frac{d}{n}\log\left(\frac{d}{n}\right)\right]\geq 1-\frac{\epsilon}{2} \tag{3.8}$$

*nets*

$$\begin{aligned}
&\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(X_i^T\hat{\theta}^{(sq)}\leq 0, X_i^T\theta_0>0\right)\leq C\frac{d}{n}\log\frac{d}{n}+\sqrt{\frac{2}{\epsilon}}\sqrt{\frac{1}{n}}\right]\\
\geq\ &\mathbb{P}\Bigg[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(X_i^T\hat{\theta}^{(sq)}\leq 0, X_i^T\theta_0>0\right)\leq a^+\|\hat{\theta}^{(sq)}-\theta_0\|_2+\sqrt{\frac{2}{\epsilon}}\sqrt{\frac{1}{n}}\ \textit{and}\\
&a^+\|\hat{\theta}^{(sq)}-\theta_0\|_2\leq C\frac{d}{n}\log\frac{d}{n}\Bigg]\\
\geq\ &1-\epsilon,
\end{aligned} \tag{3.9}$$

*thus demonstrating that the second term of (3.7) is $O_p\left(\left(\frac{d}{n}\log\frac{d}{n}\right)\vee\frac{1}{\sqrt{n}}\right)$.*

*Similarly the numerator*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(Y_i=1, X_i^T\hat{\theta}^{(sq)}\leq 0\right)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(Y_i=1, X_i^T\theta_0\leq 0\right) \tag{3.10}$$

$$+\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(Y_i=1, X_i^T\hat{\theta}^{(sq)}\leq 0, X_i^T\theta_0>0\right) \tag{3.11}$$

*equals* $P(Y=1, X_i^T\theta_0 \leq 0) + O_p\left(\frac{1}{\sqrt{n}} \vee \left(\frac{d}{n}\log\left(\frac{n}{d}\right)\right)\right)$. *Therefore* $\hat{\alpha}$ *equals*

$$\begin{aligned} & \frac{P(Y=1, X_i^T\theta_0 \leq 0) + O_p\left(\frac{1}{\sqrt{n}} \vee \left(\frac{d}{n}\log\left(\frac{n}{d}\right)\right)\right)}{P(X^T\theta_0 \leq 0) + o_p(1)} \\ = & \frac{P(Y=1, X_i^T\theta_0 \leq 0)}{P(X^T\theta_0 \leq 0)} + O_p\left(\frac{1}{\sqrt{n}} \vee \left(\frac{d}{n}\log\left(\frac{n}{d}\right)\right)\right) \\ = & \alpha_0 + O_p\left(\frac{1}{\sqrt{n}} \vee \left(\frac{d}{n}\log\left(\frac{n}{d}\right)\right)\right) \end{aligned} \tag{3.12}$$

*This is a faster convergence rate than* $\hat{\alpha}^{(sq)} - \alpha_0 = O_p\left(\sqrt{\frac{d}{n}\log\left(\frac{n}{d}\right)}\right)$. *In the case that* $d$ *grows very slowly with* $n$, *including the case where* $d = O(N^\delta)$ *for any* $\delta < 1/2$, *the rate of convergence would be the optimal rate* $\frac{1}{\sqrt{n}}$.

To put this rate into perspective, we will compare this rate of convergence with the minimax rate for all possible estimators. The minimax rate can be obtained using an application of Assouad's Lemma.

**Theorem III.3.** *Let* $\mathcal{P} = \mathcal{P}_n$ *be the set of all distributions of* $(X, Y)$ *that satisfy the model in the beginning of Section 3.2, Assumptions A1 and A2 are satisfied, and the parameters of* $\alpha_0$ *and* $\beta_0$ *are known. Then we have the min-max lower bound*

$$\inf_{\hat{\theta}} \sup_{\theta_0 = \theta \in S^{d-1}} \mathbb{E}\left(\|\hat{\theta} - \theta\|^2\right) \gtrsim \left(\frac{d}{n}\right)^2 \tag{3.13}$$

*with the infimum taken over all possible estimators of* $\theta_0$.

*Proof.* See Section B.0.0.4. □

The least squares estimator provides a starting route to analyze the model in Section 3.2, as it is consistent at a near min-max rate. Unfortunately, there are methodological issues to implementation, namely that the objective function is not continuous. This shortcoming becomes more severe due to our consideration of a

model where the dimension can grow with $n$. Resolving this issue will require some indirect procedures, and in later sections we will direct our analysis to the optimization of manageable surrogate expressions.

## 3.4 Alternative Estimator

The least squares estimator exhibits desirable convergence, but it has the downside of requiring the estimation of $\alpha_0$ and $\beta_0$ to be concurrent with the estimation of $\theta_0$, when only the latter parameter is our main interest. To simplify estimation, we investigate whether there exists a method to estimate $\theta_0$ that is decoupled from the estimation of $\alpha_0$ and $\beta_0$.

Before we discuss the estimator that results from this investigation, we first provide a train of thought as motivation. We first simplify the least squares loss function, when $\alpha$ and $\beta$ are set to their true values of $\alpha_0$ and $\beta_0$:

$$\begin{aligned}
& \left[Y - \alpha_0 \cdot \mathbb{1}(X^T\theta \leq 0) - \beta_0 \cdot \mathbb{1}(X^T\theta > 0)\right]^2 \\
= \;& Y^2 + \left(\alpha_0^2 - 2\alpha_0 Y\right)\mathbb{1}(X^T\theta \leq 0) + \left(\beta_0^2 - 2\beta_0 Y\right)\mathbb{1}(X^T\theta > 0) \\
= \;& Y^2 + \left(\beta_0^2 - 2\beta_0 Y\right) + \left(\alpha_0^2 - \beta_0^2 - 2(\alpha_0 - \beta_0)Y\right)\mathbb{1}(X^T\theta \leq 0) \\
= \;& Y^2 + \left(\beta_0^2 - 2\beta_0 Y\right) + 2(\beta_0 - \alpha_0)\left(Y - \frac{\beta_0 + \alpha_0}{2}\right)\mathbb{1}(X^T\theta \leq 0). \qquad (3.14)
\end{aligned}$$

This equivalence means that minimizing the least squares expression ($\mathbb{P}_n$ applied to the first line), over possible values of $\theta$ is equivalent to minimizing the last line. Since the only term in the last line that depends on $\theta$ is the third term, this would be the same as minimizing the expression

$$\mathbb{P}_n\left(Y - \frac{\beta_0 + \alpha_0}{2}\right)\mathbb{1}(X^T\theta \leq 0). \qquad (3.15)$$

Although this term still involves $\alpha_0$ and $\beta_0$ (it is not the signal independent estimator that we want), it turns out that the $\frac{\alpha_0+\beta_0}{2}$ term is not strictly necessary, and given Assumption A1, with $\gamma$ as any value in $(\alpha_0, \beta_0)$, the expression

$$\mathbb{E}\left[(Y-\gamma)\cdot \mathbb{1}(X^T\theta \le 0)\right] \tag{3.16}$$

is uniquely minimized when $\theta = \theta_0$ (details in Section B.0.1). This motivates estimation through minimization of $\mathbb{P}_n\left[(Y-\gamma)\mathbb{1}(X^T\theta \le 0)\right]$, so long as a $\gamma$ value can be obtained. Fortunately, the marginal expectation of $Y$ equals

$$\mathbb{E}[Y] = \mathbb{E}[Y|X] = \alpha_0\mathbb{P}[X^T\theta_0 \le 0] + \beta_0(1-\mathbb{P}[X^T\theta_0 \le 0]), \tag{3.17}$$

which is a value strictly in between $\alpha_0$ and $\beta_0$ for each $n$. This points towards replacing $\gamma$ by $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, and setting the estimator of $\theta_0$ to be

$$\hat{\theta} := \underset{\substack{\theta\in\mathbb{R}^d \\ \|\theta\|_2=1}}{\arg\min}\, \mathbb{P}_n(Y-\bar{Y})\mathbb{1}(X^T\theta \le 0). \tag{3.18}$$

This alternative estimator can be thought of as a slight modification to the least squares estimator, and therefore is expected to also be consistent and converge at the same rate to $\theta_0$. Conveniently, the consistency of the estimator can be established similarly, by using Assumption A1 to bound the value of

$$P(Y-\mathbb{E}Y)\mathbb{1}\left(X^T\theta \le 0\right) + P(\bar{Y}-\mathbb{E}Y)\mathbb{1}\left(X^T\theta \le 0\right) \tag{3.19}$$

for $\theta$ within a small neighborhood of $\theta_0$, and using a VC dimension argument alongside of Assumption A2 to obtain a Glivenko-Cantelli type result.

**Theorem III.4.** *Under Assumptions A1 and A2, $\|\hat{\theta}-\theta_0\|_2 \to 0$ in probability.*

*Proof.* See Section B.0.1.1. □

The rate of convergence can also be shown through similar steps as for the least squares estimator, since we would still be working with a bounded class of functions with VC dimension bounded by a scalar times $d$. As a result of the similarity in method, among other details left in the proof sections, the rate of convergence obtained for this alternate estimator is the same as that of the least squares estimator.

**Theorem III.5.** *Under Assumptions A1 and A2,* $\|\hat{\theta} - \theta_0\|_2 = O_p\left(\frac{d}{n}\log\left(\frac{n}{d}\right)\right)$.

*Proof.* See Section B.0.1.2. □

Due to its simplicity of not requiring the concurrent estimation of $\alpha_0$ and $\beta_0$, we focus on the former estimator in later sections.

## 3.5 Higher Dimensions with $\ell_0$ Penalty

In this section we consider the case where the dimension of the problem, $d$, is not $o(n)$. Instead, consider the case where $\liminf \frac{d}{n}$ is bounded below by a positive constant, which in particular allows $\frac{d}{n} \to \infty$. For estimation of $\theta_0$, we again consider minimization of the expression

$$\mathbb{M}_{n,d}(\theta) := \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})\mathbb{1}(X_i^T\theta \leq 0) \tag{3.20}$$

In low dimensions ($d = o(n)$), minimization was done over the entire unit sphere of $\mathbb{R}^d$ without much issue. If we consider the case where $d$ can be greater than $n$, then such a domain would be in-feasibly large. To put the problems into the realm where we can obtain results, we will consider the case where the support of the parameter $\theta_0$ (the number of nonzero entries of $\theta_0$) is much smaller than $n$. For estimation, we will pick estimators from the subset of $S^{d-1}$ where the support is no greater than $d_0$, where $d_0 = d_{0,n}$ is a sequence of positive integers satisfying the following

**Assumption A3** The dimension $d$, true support size $|\text{supp}(\theta_0)| = |\{1 \leq i \leq d : \theta_{0,i} \neq 0\}|$, and $d_0$ satisfy $d_0 \geq |\text{supp}(\theta_0)|$ and $d_0 \log\left(\frac{d}{|\text{supp}(\theta_0)|}\right) = o(n)$.

More specifically, we will be choosing our estimator from the set $\mathcal{H} := \{\theta \in S^{d-1} : \|\theta\|_0 \leq d_0\}$, and the estimator will satisfy

$$\hat{\theta}_{\mathcal{H}} := \arg\min_{\theta \in \mathcal{H}} \mathbb{M}_{n,d}(\theta) \tag{3.21}$$

As in previous sections, we wish to show that $\hat{\theta}_{\mathcal{H}}$ is consistent for $\theta_0$ and determine the rate at which it converges to the true parameter. Additionally, as in other problems in high dimensional literature, we wish to show that $\hat{\theta}_{\mathcal{H}}$ can capture the true support set with probability going to 1. In the following narrative, we show the latter before giving results for the former.

Our first result will demonstrate that given any two subsets $m_1$ and $m_2$ of $\{1, \ldots, d\}$ of size $d_0$ each, if $m_2$ contains the true support set while $m_1$ does not, then with high probability, the empirical loss function will be lower when minimizing over the estimating space of norm-1 vectors with support $m_2$ than when minimizing over the estimating space of vectors with support $m_1$.

**Theorem III.6.** *Suppose Assumptions A1 and A3 are satisfied. Let $m_1$ and $m_2$ be subsets of $\{1\ldots, d\}$ with $d_0$ elements each, such that supp$(\theta_0) \subseteq m_2$ and supp$(\theta_0) \nsubseteq m_1$. Denote*

$$\hat{\theta}_{m_j} := \arg\min_{\substack{\theta \in S^{d-1} \\ \theta_k = 0 \text{ for } k \notin m_j}} \mathbb{M}_{n,d}(\theta) \tag{3.22}$$

*for $j = 1, 2$.*

*There exists a constant $C > 0$ such that if the following inequality is satisfied:*

$$\theta_{0,min} := \min_{j:|\theta_{0,j}|\neq 0} |\theta_{0,j}| \geq C\sqrt{\frac{d_0 \log\left(\frac{d}{d_0}\right)}{n}}, \tag{3.23}$$

*then with probability at least* $1 - 2\exp\left(-Ld_0 \log\left(\frac{d}{d_0}\right)\right)$,

$$\mathbb{M}_{n,d}(\hat{\theta}_{m_2}) < \mathbb{M}_{n,d}(\hat{\theta}_{m_1}) \tag{3.24}$$

*for some constant $L > 2$ not dependent on $n$.*

*Proof.* See Section B.0.2. □

**Remark 17.** *The probability bound on the above event,* $1 - 2\exp\left(-Ld_0 \log\left(\frac{d}{d_0}\right)\right)$, *does in fact go to 1 under the high dimensional setting that we are working under. We either have the case where the support size of $\theta_0$ goes to $\infty$, meaning that $d_0 \to \infty$, $d_0 \log\left(\frac{d}{d_0}\right) \to \infty$ and the probability bound goes to 1, or the case where $d_0$ stays bounded from above, in which case $d/d_0 \to \infty$ and the probability bound also goes to 1.*

The proof of the preceding result utilizes probability inequalities for the difference between the empirical loss function and the risk function, which use the fact that the loss and risk functions have finite bounds. Combined with a bound on the difference between the risk function over the $m_1$ parameter space and the risk function over the $m_2$ parameter space, derived from the $\theta_{min}$ condition, we obtain the above theorem as a result.

A logical extension of Theorem III.6 leads to the result that with high probability, the support set of the estimator will contain the true support set. This can be done by summing the probability bound obtained in the preceding theorem over all sets of size $d_0$.

**Theorem III.7.** *Suppose Assumptions A1 and A3 are satisfied.*

*There exists a constant $C > 0$ such that if*

$$\theta_{0,min} := \min_{j:|\theta_{0,j}|\neq 0} |\theta_{0,j}| \geq C\sqrt{\frac{d_0 \log\left(\frac{d}{d_0}\right)}{n}}, \tag{3.25}$$

*then with probability greater than $1-2\exp\left(-Kd_0\log\left(\frac{d}{d_0}\right)\right)$ for some constant $K > 0$ not dependent on $n$, $supp(\hat{\theta}_{\mathcal{H}})$ will contain the support set of $\theta_0$.*

*Proof.* The probability that $supp(\theta_0) \nsubseteq supp(\hat{\theta}_{\mathcal{H}})$ would be bounded by the sum of the probabilities that $supp(\hat{\theta}_{\mathcal{H}}) = m$ over all possible $m \subseteq \{1,\ldots,d\}$ where $|m| = d_0$ and $supp(\theta_0) \nsubseteq m$. If $m_2$ is any size $d_0$ subset which does contain $supp(\theta_0)$, then the aforementioned probability is bounded by

$$\begin{aligned}
& \sum_{supp(\theta_0)\nsubseteq m;|m|=d_0} \mathbb{P}\left[\mathbb{M}_{n,d}(\hat{\theta}_m) < \mathbb{M}_{n,d}(\hat{\theta}_{m_2})\right] \\
\leq & \sum_{supp(\theta_0)\nsubseteq m;|m|=d_0} 2\exp\left(-Ld_0\log\left(\frac{d}{d_0}\right)\right) \\
& \text{for some constant } L > 1 \text{ not dependent on } n \\
\leq & \binom{d}{d_0}\cdot 2\exp\left(-Ld_0\log\left(\frac{d}{d_0}\right)\right) \\
\leq & \left(\frac{d}{d_0}\right)^{d_0}\cdot 2\exp\left(-Ld_0\log\left(\frac{d}{d_0}\right)\right) \\
\leq & 2\exp\left(-(L-1)d_0\log\left(\frac{d}{d_0}\right)\right)
\end{aligned} \tag{3.26}$$

where the value of $L$ is the same as the $L$ appearing in Theorem III.6, which stated $L > 2$. □

As in Remark 17, note that the probability $2\exp\left(-(L-1)d_0\log\left(\frac{d}{d_0}\right)\right)$ must go to zero in the high dimensional setting.

Next, we focus on the deviation of $\hat{\theta}_{\mathcal{H}}$ from $\theta_0$ in terms of Euclidean distance. Our estimation space, $\mathcal{H}$, has a VC dimension bounded by a constant multiple of $d_0 \log\left(\frac{d}{d_0}\right)$, which is $o(n)$ by Assumption A4. Using this property, we can show the consistency and rate of convergence of the $\hat{\theta}_{\mathcal{H}}$ estimator.

**Theorem III.8.** *Suppose Assumptions A1 and A3 are satisfied, then* $\|\hat{\theta}_{\mathcal{H}} - \theta_0\|_2 \to 0$ *in probability. Specifically we have*

$$\|\hat{\theta}_{\mathcal{H}} - \theta_0\|_2 = O_p\left(\frac{d_0 \log\left(\frac{d}{d_0}\right)}{n} \log\left(d_0 \log\left(\frac{d}{d_0}\right)\right)\right) \tag{3.27}$$

## 3.6 Change Planes with Continuous Response

Our previous analysis of the change plane model, restricted to the setting where the response $Y$ is conditionally Bernoulli, allowed us to derive important asymptotic results using M-estimation techniques. The work done for the discrete response model characterized the optimal rates of convergence, in both low and high dimensional settings. Intuitively, if the conditional distribution of the response variable has fast decreasing tails, then the asymptotic behavior will be the same as in the binary response model. In this section we will perform some investigation on whether this is the case, with a result on the convergence of the least squares estimator in the low dimensional case.

As in the discrete response case, we will consider a simple model with a single hyperplane delineating two distinct distributions on either side. The two distributions will differ in their mean:

$$Y = \mu \cdot 1(X^T\theta_0 > 0) + \varepsilon \tag{3.28}$$

where $\theta_0 = \theta_{0,n} \in S^{d-1}$ is a parameter of Euclidean norm 1, $\mu > 0$ is a positive constant, $X$ is a random variable in $\mathbb{R}^d$, and $\varepsilon$ is a zero-mean, variance $\sigma^2 < \infty$ error term independent of $X$. We will further assume that $\mu$ is a known value, as generally we can take the set of $\{Y_i\}$ as a data from a mixture distribution and use an EM algorithm to estimate for the two means.

As in the discrete response model, we begin our analysis using the least squares estimator. Let

$$\begin{aligned} \hat{\theta}^{(sq)} &:= \arg\min_{\theta} \mathbb{P}_n \left[ \left( Y - \mu 1(X^T\theta > 0) \right)^2 \right] \\ &= \arg\max_{\theta} \mathbb{P}_n \left[ \left( Y - \frac{\mu}{2} \right) 1(X^T\theta > 0) \right]. \end{aligned} \tag{3.29}$$

Due to the similarities between this model and the binary response model, it is expected that with similar assumptions, the convergence rate of this estimator would be similar. We impose the same assumptions as the discrete model:

**Assumption B1** There exist positive constants $a^-$, $a^+$ such that for every $n$,

$$a^- \|\theta_1 - \theta_2\|_2 \leq P(X^T\theta_1 < 0 < X^T\theta_2) \leq a^+ \|\theta_1 - \theta_2\|_2 \tag{3.30}$$

for any $\theta_1, \theta_2 \in S^{d-1}$.

**Assumption B2** The dimension $d$ satisfies $d/n \to 0$.

Assumption B1 again allows us to work in well-behaved settings and rule out pathological instances where the distribution of $X$ would not allow $\theta_0$ to be estimable. Assumption B2 allows the use of VC dimension arguments to derive our results. With these assumptions we can show that the least squares estimators are consistent:

**Theorem III.9.** *Suppose Assumptions B1 and B2 are satisfied, then* $\|\hat{\theta}^{(sq)} - \theta_0\|_2 \xrightarrow{p} 0$

*Proof.* See Section B.0.3. □

Additionally, the least square estimator has same rate of convergence as in the binary response model:

**Theorem III.10.** *Suppose Assumptions B1 and B2 are satisfied, then* $\|\hat{\theta}^{(sq)} - \theta_0\|_2 = O_p\left(\frac{d}{n}\log\left(\frac{n}{d}\right)\right)$.

*Proof.* See Section B.0.4. □

### 3.6.1 Future Work

We have examined and derived a variety of results for the change plane problem. For the discrete response model, we were able to derive asymptotic properties for both the low dimensional and high dimensional cases. Specifically, we derived rates of convergence in both cases, and demonstrated support recovery properties specific to the high dimensional case. For the continuous response model, our current results is limited to the low dimensional case. Naturally, we would like to give the same treatment to the continuous response model, and derive asymptotic results for the high dimensional case. To do this would require a different strategy, due to the fact that the high dimensional results for the discrete response model made use of the McDiarmid inequality, which unfortunately is not known to generally extend to continuous variables.

An important methodological issue in the change plane problem is computation, which involves the optimization of non-continuous expressions as one can see from expressions (3.3), (3.18), and (3.29). This non-standard minimization problem is made more difficult due to working with high dimensional parameters, which means that unlike the one dimensional change point problem, the problem cannot be done

by searching through a reasonable number of candidate values in the same way the change point problem only requires searching through all points along a linear line. To make the optimization problem more feasible, we will focus on a strategy employed in *Seo and Linton* (2007) and *Li et al.* (2018): using a smoothed objective function in the place of the true objective function.

Consider the binary response model in the $d = o(n)$ case. take a kernel function $k : \mathbb{R} \to \mathbb{R}$ which is non-negative, smooth, and has an integral of 1. Letting $K(x) := \int_{-\infty}^{x} k(y)\, dy$, we have a function which is positive, smooth, non-decreasing, and has limits of 0 and 1 at $-\infty$ and $+\infty$, respectively. Furthermore, letting $\xi$ be a positive value, the function $K(\xi x)$ will converge pointwise to the indicator function $1(x > 0)$. Therefore, if we denote

$$\hat{\theta}^{(\xi)} := \underset{\theta \in \mathbb{R}^d;\ \|\theta\|_2 = 1}{\arg\min}\ \mathbb{P}_n (Y - \bar{Y}) K(-\xi X^T \theta), \tag{3.31}$$

then $\hat{\theta}^{(\xi)}$ could serve as a feasibly computable proxy to the estimator $\hat{\theta}$. It is reasonable to predict that under some regularity conditions, and if $\xi = \xi_n$ is some increasing sequence, $\hat{\theta}^{(\xi)}$ should be a consistent estimator for the true parameter $\theta_0$. Its rate of convergence will be different depending on how fast $\xi_n$ increases with $n$, and such details can be explored in detail in future work.

Alternatively, another way to convert the non-continuous problem to a smooth one would be to compare it to a logistic problem. Consider the continuous response model with $\mu = 1$: the mean response will be the function $1(X^T \theta_0 > 0)$, which is approximately the compressed logistic curve $1/(1 + \exp(-\xi X^T \theta_0))$ for a very large $\xi$. This motivates using an expression similar to the log-likelihood expression for logistic

regression:

$$\hat{\theta}^{\xi} = \underset{\theta \in \mathbb{R}^d}{\arg\min} \mathbb{P}_n \left[ \frac{1}{\xi} \log\left(1 + \exp(\xi X^T \theta)\right) - Y X^T \theta \right]. \tag{3.32}$$

This alternate expression could not be compared to the objective functions we have analyzed earlier, but it does have the enticing property of being a convex expression over $d \in \mathbb{R}^d$. It may be useful to consider this second approach in the high dimensional setting by adding an $\ell_1$ penalizer, as the sum would remain a convex expression.

Overall there still remains work to be done in this project. In the continuous response model, it is desirable to obtain similar high dimensional results. Additionally, the strategy of utilizing smooth proxy expressions is also an open area with multiple problems to solve.

# APPENDICES

# APPENDIX A

# Proofs and Extra Analysis for Chapter II

## A.1 Analysis and Proofs for Single Change Point Problem

### A.1.1 Problem Setup

Instead of proving Theorems II.1 and II.2 directly, we shall consider a more general nonparametric result from which the two theorems will follow as a special case. As before suppose the time series data is $(x_1, Y_1), ..., (x_N, Y_N)$, where $x_i = i/N$ and $Y_i = f(x_i) + \varepsilon_i$ for $i = 1, ..., N$. We will make the weaker assumptions that

- $f$ is a right continuous function in $[0, 1]$ with a single left discontinuity at some point $\tau_0 \in (0, 1)$, with jump size $f(\tau_0+) - f(\tau_0-) = \Delta$
- there exists a $\beta_f > 0$ where $|f(x)-f(y)| \leq \beta_f|x-y|$ whenever $(x-\tau_0)(y-\tau_0) > 0$
- the errors $\varepsilon_i$'s are iid $N(0, \sigma^2)$ error terms

The main difference between this model and the model presented in section 2.2.1 is the looser restriction on the signal $f$: here $f$ could be any Lipschitz continuous function with a single discontinuity and not be constrained to the family of piecewise constant functions.

We will first remark on some background regarding this this model before moving on to proving some results. Estimation procedures for such a dataset can be found in *Loader et al.* (1996), where one-sided polynomial fitting was used to obtain an estimate $\hat{\tau}_N$ for $\tau_N := \lfloor N\tau_0 \rfloor / N$. In summary, fix a sequence of bandwidth $h = h_N$, a non-negative integer $p$, and a kernel function $K$ with support on $[-1, 1]$. Next, for all $x_m \in (h, 1-h)$, consider the signal estimates

$$\begin{aligned}
\hat{f}_-(x_m) &:= \pi_1 \left( \underset{(a_0,\ldots,a_p)\in\mathbb{R}^{p+1}}{\arg\min} \left( \sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right) \left(Y_{m-j-1} - a_0 - a_1 j - \ldots - a_p j^p\right)^2 \right) \right) \\
\hat{f}_+(x_m) &:= \pi_1 \left( \underset{(a_0,\ldots,a_p)\in\mathbb{R}^{p+1}}{\arg\min} \left( \sum_{j=0}^{Nh} K\left(\frac{j}{nh}\right) \left(Y_{m+j} - a_0 - a_1 j - \ldots - a_p j^p\right)^2 \right) \right),
\end{aligned} \tag{A.1}$$

where $\pi_1$ is the projection functions such that $\pi_1(a_0, \ldots, a_p) = a_0$. The change point estimate is

$$\hat{\tau}_N := \underset{x_i\in(h,1-h)}{\arg\max} |\hat{f}_+(x_i) - \hat{f}_-(x_i)|. \tag{A.2}$$

This estimator is consistent under a few regularity conditions on the kernel $K$ and conditions on how fast $h$ converges to 0. For the sake of brevity we will not mention all those conditions here, but we will note that under said conditions, $\mathbb{P}[N(\hat{\tau}_N - \tau_N) = k] \to \mathbb{P}[L(\Delta/\sigma) = k]$ for all $k \in \mathbb{Z}$, which is the exactly asymptotic result for $\hat{\tau}_N$ obtained by least squares in a stump model setting. Finally, for our purposes we propose estimators $\hat{\alpha}$ and $\hat{\beta}$ for $f(\tau_0-)$ and $f(\tau_0+)$, respectively, defining them as

$$\begin{aligned}
\hat{\alpha} &:= \frac{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right) Y_{N\hat{\tau}_N - j - 1}}{\sum_{j=1}^{Nh} K\left(\frac{j}{Nh}\right)} \\
\hat{\beta} &:= \frac{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right) Y_{N\hat{\tau}_N + j}}{\sum_{j=1}^{Nh} K\left(\frac{j}{Nh}\right)}.
\end{aligned} \tag{A.3}$$

These two estimators are consistent:

**Lemma 2.** *$|\hat{\alpha} - f(\tau_0+)|$ and $|\hat{\beta} - f(\tau_0-)|$ are $O_p(h \vee (Nh)^{-1/2})$.*

It is possible to perform intelligent sampling to this nonparametric setting as in steps (ISS1)-(ISS4), though with a slight adjustment. Instead of fitting a stump function at step (ISS2), use one-sided local polynomial fitting with bandwidth $h$ on the first stage subsample to obtain estimates $(\hat{\alpha}^{(1)}, \hat{\beta}^{(1)}, \hat{\tau}_N)$ for the parameters $(f(\tau_0-), f(\tau_0+), \tau_N)$. These first stage estimators satisfy the following consistency result:

$$\mathbb{P}\left[|\hat{\tau}_N^{(1)} - \tau_N| \leq w(N); \quad |\hat{\alpha}^{(1)} - f(\tau_0-)| \vee |\hat{\beta}^{(1)} - f(\tau_0+)| \leq \rho_N\right] \to 1 \tag{A.4}$$

for the sequence $w(N) = CN^{1-\gamma+\delta}$ where $\delta$ and $C$ can be any positive constants, and some sequence $\rho_N \to 0$ (an explicit sequence can be derived by Lemma 2). The consistency condition in (A.4) is sufficient for a generalized versions of Theorems II.1 and II.2:

**Theorem A.1.**

$$\hat{\tau}^{(2)} - \tau_N = O_p(N^{-1}) \tag{A.5}$$

**Theorem A.2.** *Suppose the conditions of Theorem A.2 are satisfied, then for all integers $k \in \mathbb{Z}$ we have*

$$\mathbb{P}\left[\lambda_2\left(\tau_N, \hat{\tau}_N^{(2)}\right) = k\right] \to \mathbb{P}\left[L_{\Delta/\sigma} = k\right] \tag{A.6}$$

These results can hold under a general nonparametric setting, but they still hold for the stump model from Section 2.2. Since consistency condition (A.4) is if $f$ is a stump function and least square fitting was used at step (ISS2) as it was written in

Section 2.2.2, Theorems A.1 and A.2 do imply Theorems II.1 and II.2. The proof of Theorem A.1 will be covered in Section A.1.3, while the proof of Theorem A.2 will be covered in Appendix B at Section A.1.4.

**Remark 18.** *We note that not only do the consistency results of intelligent sampling for stump models generalize to this nonparametric setting, the computational time aspects translates also does not change. Local polynomial fitting on n data points takes $O(N)$ computational time, see e.g. Seifert et al. (1994). Therefore the computational time analysis in Section 2.2.1 still holds for this nonparametric case.*

**Remark 19.** *It is not possible derive asymptotic distribution results for $N(\hat{\tau}_N^{(2)} - \tau_N)$. Consider the case where $\tau = 0.5$, $N_1 = \sqrt{N}$ (or $\gamma = 0.5$), and the two subsequences $N = 2^{2j}$ or $N = 3^{2j}$ for some large integer $j$. In such cases the first stage subsample would choose points that have integer multiples of $1/2^j$ or $1/3^j$ as their x-coordinate.*

- *If $N = 2^{2j}$, $\tau_N = \frac{\lfloor 2^{2j} \cdot 0.5 \rfloor}{2^{2j}} = 0.5$ is an integer multiple of $1/2^j$, and hence $\tau_N$ is an x-coordinate used in the first stage.*
- *If $N = 3^{2j}$, then $\tau_N = \frac{\lfloor 3^{2j} \cdot 0.5 \rfloor}{3^{2j}}$, and it can be checked that $\lfloor 3^{2j} \cdot 0.5 \rfloor$ is an even integer not divisible by 3. Since the x-coordinate of every first stage data point takes the form $\frac{k}{3^j} = \frac{3^j k}{3^{2j}}$ for some integer $k$, this means $\tau_N$ is not used in the first stage.*

*Hence, in the former case we cannot ever have $\hat{\tau}_N^{(2)} = \tau_N$, while in the latter case, we have $\tau_N^{(2)} = \tau_N$ and Theorem II.2 tells us that $\mathbb{P}[\hat{\tau}_N^{(2)} = \tau_N]$ converges to the nonzero $\mathbb{P}[L(\Delta/\sigma) = 0]$ as $j$ increases. Clearly, we have two subsequences for which $\mathbb{P}[\hat{\tau}_N^{(2)} = \tau_N]$ converges to different values.*

To further validate the extension to this nonparametric setting, we also ran a set of simulations for when $Y_i = 2\sin(4\pi x_i) + 2 \cdot 1(x_i > 0.5) + \varepsilon_i$, where $\varepsilon_i$ are iid $N(0, 1)$. We took 15 values of $N$ between 2500 and $10^6$, chosen evenly on the log

scale, and applied intelligent sampling on 1000 replicates. For each of these values of $N$. First stage used roughly $N_1 = \sqrt{N}$ points, which were subjected to one sided local polynomial fitting with a parabolic kernel and bandwidth $h = N_1^{-0.3}$, while the second stage interval had half-width $8/\sqrt{N}$. Figures A.1 and A.2 show results consistent with Theorems A.1 and A.2.

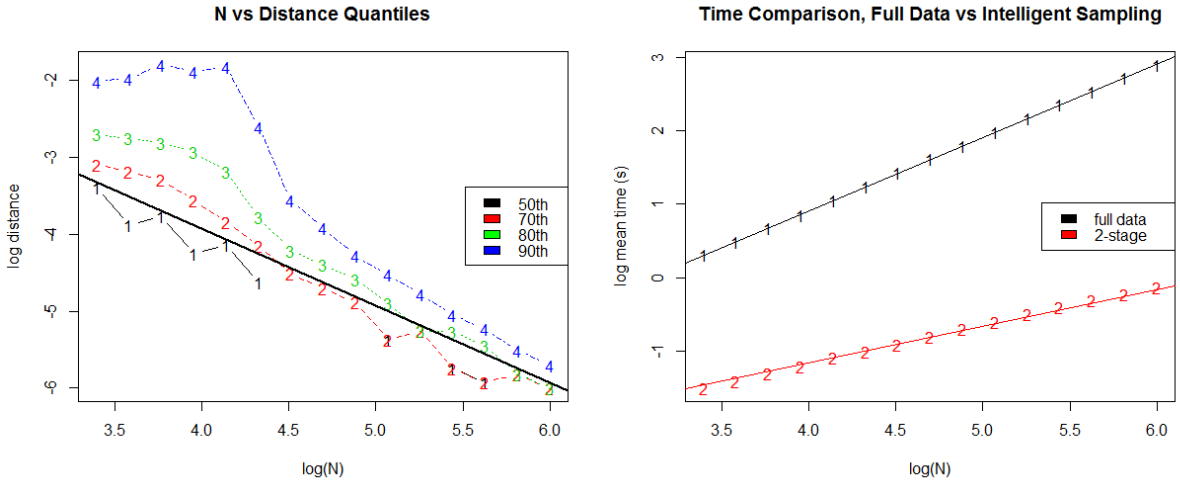


Figure A.1: left graph shows log-log plot of the quantiles of $|\hat{\tau}_N^{(2)} - \tau_N|$ versus $N$, with the solid black line having a slope of exactly -1. Some datapoints for the quantiles of the 50th quantiles do not appear since for some $N$, the median of $|\hat{\tau}_N^{(2)} - \tau_N|$ was 0. Right graph is a log-log plot of the mean computational time of using all datapoints (black) and intelligent sampling (red), with the solid black line having a slope of exactly 1 and the solid red a slope of exactly 0.5.

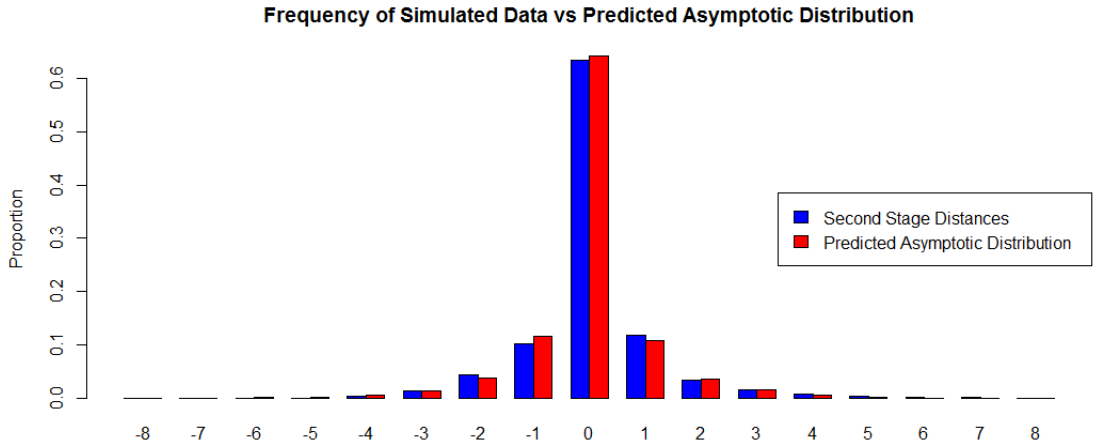


Figure A.2: distribution of $\lambda_2\left(\tau_N, \hat{\tau}^{(2)}\right)$ values (blue) compared with the distribution of $L$ from Theorem II.2.

### A.1.2 Proof of Corollary 2

*Proof.* We will show that for any $\epsilon > 0$,

$$\mathbb{P}\left[|\hat{\beta} - f(\tau+)| > C_0(h \vee (Nh)^{-1/2})\right] \leq \epsilon. \tag{A.7}$$

We start off by utilizing the rate of convergence of the change point estimator: there is a constant $C_1 > 0$ such that

$$\mathbb{P}\left[|\hat{\tau}_N - \tau| > \frac{C_1}{N}\right] < \frac{\epsilon}{2}$$

for all sufficiently large $N$. Hence, for any $C > 0$ we have,

$$\begin{aligned}
&\mathbb{P}\Big[|\hat{\beta}_N) - f(\tau+)| > C\Big] \leq \\
&\mathbb{P}\left[|\hat{\beta}_N - f(\tau+)| > C \text{ and } |\hat{\tau}_N - \tau| \leq \frac{C_1}{N}\right] + \mathbb{P}\left[|\hat{\tau}_N - \tau| > \frac{C_1}{N}\right] \leq \\
&\mathbb{P}\left[|\hat{\beta}_N(t) - f(\tau+)| > C \text{ for some } |t - \tau| \leq \frac{C_1}{N}\right] + \frac{\epsilon}{2} \qquad \text{(A.8)}
\end{aligned}$$

where

$$\hat{\beta}_N(t) := \frac{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right) Y_{Nt+j}}{\sum_{j=1}^{Nh} K\left(\frac{j}{Nh}\right)} \qquad \text{(A.9)}$$

Next, we bound the first term above, so consider only the case where $|t - \tau| \leq \frac{C_1}{N}$. By expanding we have

$$\begin{aligned}
|\hat{\beta}(t) - f(\tau+)| &= \left|\frac{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)[f(t + j/N) + \varepsilon_{Nt+j}]}{\sum_{j=1}^{Nh} K\left(\frac{j}{Nh}\right)} - f(\tau+)\right| \\
&\leq \frac{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)|f(t + j/N) - f(\tau+)|}{\sum_{j=1}^{Nh} K\left(\frac{j}{Nh}\right)} + \left|\frac{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)\varepsilon_{Nt+j}}{\sum_{j=1}^{Nh} K\left(\frac{j}{Nh}\right)}\right| \\
&:= A(t) + |B(t)| \qquad \text{(A.10)}
\end{aligned}$$

First, we derive a bound for $A(t)$. If $t \geq \tau$ then we have

$$\begin{aligned}
A(t) &= \frac{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)|f(t + j/N) - f(\tau+)|}{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)} \\
&\leq \frac{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)\beta_f|t + j/N - \tau|}{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)} \\
&\leq \frac{\beta_f \sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)\left(\frac{C_1+j}{N}\right)}{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)} \\
&= \beta_f h \frac{\frac{1}{Nh}\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)\left(\frac{j}{Nh}\right)}{\frac{1}{Nh}\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)} + \frac{C_1}{N} \qquad \text{(A.11)}
\end{aligned}$$

Note that since $Nh \to \infty$ as $N \to \infty$, we have $\frac{1}{Nh}\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)\frac{j}{Nh} \to \int_0^1 xK(x)\,dx$ (which exists) and $\sum_{j=1}^{Nh} K\left(\frac{j}{Nh}\right) \to \int_0^1 K(x)\,dx = 1$, as $N \to \infty$, hence we can find a constant $M > 0$ such that

$$A(t) \le \beta_f Mh + \frac{C_1}{N} \tag{A.12}$$

for all sufficiently large $N$. On the other hand, suppose $t < \tau$. For sufficiently large $N$ we would have $N(\tau - t) \le C_1 < Nh$ and so

$$\begin{aligned}
A(t) &= \frac{\sum_{j=0}^{N(\tau-t)-1} K\left(\frac{j}{Nh}\right)|f(t+j/N) - f(\tau+)|}{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)} + \\
&\quad \frac{\sum_{j=N(\tau-t)}^{Nh} K\left(\frac{j}{Nh}\right)|f(t+j/N) - f(\tau+)|}{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)} \\
&\le \frac{\sum_{j=0}^{N(\tau-t)-1} K\left(\frac{j}{Nh}\right)(\Delta + \beta_f(\tau - t - j/N))}{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)} + \\
&\quad \frac{\sum_{j=N(\tau-t)}^{Nh} K\left(\frac{j}{Nh}\right)\beta_f(t + j/N - \tau)}{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)} \\
&\le \frac{K^{\uparrow} N(\tau - t)\Delta}{\sum_{j=1}^{Nh} K\left(\frac{j}{Nh}\right)} + \frac{\beta_f \sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)\left(\frac{C_1+j}{N}\right)}{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)} \\
&\quad (K^{\uparrow} \text{ is any constants that uniformly bounds the function } K \text{ from above on} [0,1]) \\
&\le (K^{\uparrow}\Delta)\frac{\frac{1}{Nh}C_1}{\frac{1}{Nh}\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)} + \beta_f h \frac{\frac{1}{Nh}\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)\left(\frac{j}{Nh}\right)}{\frac{1}{Nh}\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)} + \frac{C_1}{N}
\end{aligned} \tag{A.13}$$

which, for sufficiently large $N$, can be bounded by $\frac{M_1}{Nh} + \beta_f Mh + \frac{C_1}{N}$ for some constants $M, M_1 > 0$. Hence, this shows that $A(t)$ itself is $O(h \vee (Nh)^{-1})$ for all $t$ where $|t - \tau| \le \frac{C_1}{N}$.

Next, we consider the random term

$$B(t) = \frac{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)\varepsilon_{Nt+j}}{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)} \tag{A.14}$$

which satisfies

$$\begin{aligned}
\mathbb{E}[B(t)] &= 0 \\
\mathrm{var}(B(t)) &= \frac{\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)^2}{\left(\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)\right)^2} \\
&= (Nh)^{-1} \frac{\frac{1}{Nh}\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)^2}{\left(\frac{1}{Nh}\sum_{j=0}^{Nh} K\left(\frac{j}{Nh}\right)\right)^2} \\
&\leq (Nh)^{-1} 2 \int_0^1 K(x)^2\, dx, \qquad \text{for all sufficiently large } N. \quad \text{(A.15)}
\end{aligned}$$

Thus $B(t) = O_p((Nh)^{-1/2})$ by Chebychev's inequality.

Combining these results on $A(t)$ and $B(t)$ derived above, one can find constants $C_2, C_3 > 0$ such that for all $N > N_2$ for some integer $N_2$ we have

$$\begin{aligned}
A(t) &\leq C_2[h \vee (Nh)^{-1}] \\
\mathbb{P}[|B(t)| > C_3(h \vee (Nh)^{-1/2})] &\leq \frac{\epsilon}{2(2C_1+3)} \qquad \text{(A.16)}
\end{aligned}$$

for all $|t-\tau| \leq \frac{C_1}{N}$, to get from (A.8):

$$\begin{aligned}
&\mathbb{P}\Big[|\hat{\beta}_N - f(\tau+)| > (C_1+C_2)(h \vee (Nh)^{-1/2})\Big] \leq \\
&\mathbb{P}\left[|\hat{\beta}_N(t) - f(\tau+)| > (C_1+C_2)(h \vee (Nh)^{-1/2}) \text{ for some } |t-\tau| \leq \frac{C_1}{N}\right] + \frac{\epsilon}{2} \leq \\
&\mathbb{P}\left[A(t) + |B(t)| > (C_1+C_2)(h \vee (Nh)^{-1/2}) \text{ for some } |t-\tau| \leq \frac{C_1}{N}\right] + \frac{\epsilon}{2} \leq \\
&\sum_{t:|t-\tau|\leq C_1/N} \mathbb{P}\left[A(t) + |B(t)| > (C_1+C_2)(h \vee (Nh)^{-1/2})\right] + \frac{\epsilon}{2} \leq \\
&\sum_{t:|t-\tau|\leq C_1/n} \mathbb{P}\left[|B(t)| > C_2(h \vee (Nh)^{-1/2})\right] + \frac{\epsilon}{2} \leq \\
&\sum_{t:|t-\tau|\leq C_1/N} \frac{\epsilon}{2(2C_1+3)} + \frac{\epsilon}{2} \leq \epsilon \qquad \text{(A.17)}
\end{aligned}$$

for all $N \geq N_1 \vee N_2$. This establishes that $|\hat{\beta}_N - f(\tau+)|$ is $O_p(h \vee (Nh)^{-1/2})$, and the proof for $|\hat{\alpha}_N - f(\tau-)|$ proceeds similarly. □

### A.1.3 Proof of Theorem A.1

The structure of this proof will be similar to the rate of convergence proof found in Lan et al (2007). We will initially set some notations: let $\tau_N := \lfloor N\tau \rfloor / N$, and define

$$\tau_N^{(2)} := \begin{cases} \tau_N & \text{if } \tau_N \text{ is not in first subsample} \\ \tau_N - 1/N & \text{if } \tau_N \text{ is in first subsample} \end{cases} \tag{A.18}$$

We will show that $\left(\hat{\tau}_N^{(2)} - \tau_N^{(2)}\right)$ is $O_p(1/N)$, which will also demonstrate the same rate of convergence for $\left(\hat{\tau}_N^{(2)} - \tau_N\right)$. An additional property of $\tau_N^{(2)}$, used later on, is the fact that $\lambda_2\left(\tau_N^{(2)}, \hat{\tau}_N^{(2)}\right) = \lambda_2\left(\tau_N, \hat{\tau}_N^{(2)}\right)$. This will be utilized in the proof of Theorems II.2 and II.5.

*Proof.* Denote $G_N$ as the joint distribution of $(\hat{\alpha}^{(1)}, \hat{\beta}^{(1)}, \hat{\tau}_N^{(1)})$. Given any constant $\epsilon > 0$, there is a positive constant $C_\epsilon$ such that for all sufficiently large $N$ we have

$$\begin{aligned} (\hat{\alpha}^{(1)}, \hat{\beta}^{(1)}, \hat{\tau}_N^{(1)}) \quad &\in \quad [f(\tau-) - \rho_N, f(\tau-) + \rho_N] \\ &\times \quad [f(\tau+) - \rho_N, f(\tau+) + \rho_N] \times [\tau - C_\epsilon/N^\gamma, \tau + C_\epsilon/N^\gamma] \end{aligned} \tag{A.19}$$

with probability at least $1 - \epsilon$. Denote this event as $R_N$. It follows that for any sequence $\{a_N\}$,

$$\begin{aligned} &\mathbb{P}\Big[N|\hat{\tau}_N^{(2)} - \tau_N^{(2)}| > a_N\Big] \leq \\ &\int_{R_N} \mathbb{P}\Big[N|\hat{\tau}_N^{(2)} - \tau_N^{(2)}| > a_N \Big| (\hat{\alpha}^{(1)}, \hat{\beta}^{(1)}, \hat{\tau}_N^{(1)}) = (\alpha, \beta, t)\Big]\, dG_N(\alpha, \beta, t) + \epsilon \leq \\ &\sup_{(\alpha,\beta,t) \in R_N} \mathbb{P}\Big[N|\hat{\tau}_N^{(2)} - \tau_N^{(2)}| > a_N \Big| (\hat{\alpha}^{(1)}, \hat{\beta}^{(1)}, \hat{\tau}_N^{(1)}) = (\alpha, \beta, t)\Big] + \epsilon \end{aligned} \tag{A.20}$$

Next, we show that this first term is smaller than any $\epsilon > 0$ for a sequence $a_N = O(1/N)$ and all sufficiently large $N$, by bounding the probability that

$$\mathbb{P}_{\alpha,\beta,t}\left[N|\hat{\tau}_N^{(2)} - \tau| > a_N\right] := \mathbb{P}\left[N|\hat{\tau}_N^{(2)} - \tau| > a_N \middle| (\hat{\alpha}^{(1)}, \hat{\beta}^{(1)}, \hat{\tau}_N^{(1)}) = (\alpha, \beta, t)\right]$$

for any given $(\alpha, \beta, t) \in R_N$.

Conditional on the first stage estimates equaling $(\alpha, \beta, t)$, we can rewrite $\hat{\tau}_N^{(2)}$ and $\tau$ as maximizers of:

$$\begin{aligned}
\hat{\tau}_N^{(2)} &= \underset{d \in S^{(2)}}{\arg\min} \left( \frac{1}{\lambda_2(S^{(2)}(t))} \sum_{i: i/N \in S^{(2)}(t)} \left( Y_i - \frac{\alpha+\beta}{2} \right) (1(i/N \leq d) - 1(i/N \leq \tau)) \right) \\
&:= \underset{d \in S^{(2)}}{\arg\min}\, \mathbb{M}_n(d) \\
\tau_N^{(2)} &= \underset{d \in S^{(2)}}{\arg\min} \left( \frac{1}{\lambda_2(S^{(2)}(t))} \sum_{i: i/N \in S^{(2)}(t)} \left( Y_i - \frac{\alpha+\beta}{2} \right) (1(i/N \leq d) - 1(i/N \leq \tau)) \right) \\
&:= \underset{d \in S^{(2)}}{\arg\min}\, M_n(d)
\end{aligned} \tag{A.21}$$

Since $\tau_N^{(2)} - C_\epsilon/N^\gamma \leq t \leq \tau_N^{(2)} + C_\epsilon/N^\gamma + 2/N$, for $N$ large enough so that $KN^\delta/2 > (C_\epsilon + 2)$, we have $t - KN^{-\gamma+\delta} < \tau_N^{(2)} - KN^{-\gamma+\delta}/2 < \tau_N^{(2)} + KN^{-\gamma+\delta}/2 < t + KN^{-\gamma+\delta}$. This enables us to define the function $A(r)$ in the domain where $8N^{-1+\gamma-\delta} < r < K/2$, such that

$$\begin{aligned}
a(r) &:= \min\{M_n(d) : |d - \tau_N^{(2)}| \geq rN^{-\gamma+\delta}\} \\
&= \min_{|d-\tau^{(2)}| \geq rN^{-\gamma+\delta}} \frac{\sum_{i: i/N \in S^{(2)}} \left( f(i/N) - \frac{\alpha+\beta}{2} \right) (1(i/N \leq d) - 1(i/N \leq \tau))}{\lambda_2[t - KN^{-\gamma+\delta}, t + KN^{-\gamma+\delta}]}
\end{aligned} \tag{A.22}$$

To make $a(r)$ simpler to work with, we show that for sufficiently large $N$, there exists

a constant $A > 0$ such that $a(r) \geq Ar$. First, because

$$\begin{aligned} S^{(2)}(t) &:= \left\{ i/N : i \quad \in \mathbb{N}, \quad i/N \in [t \pm KN^{-\gamma+\delta}], \quad \lfloor N/N_1 \rfloor \text{ does not evenly divide } i \right\} \\ &\subset [t - KN^{-\gamma+\delta}, t + KN^{-\gamma+\delta}] \\ &\subset [\tau^{(2)} - 2KN^{-\gamma+\delta}, \tau^{(2)} + 2KN^{-\gamma+\delta}] \end{aligned} \tag{A.23}$$

this implies

$$\begin{aligned} |f(i/N) - f(\tau+)| &\leq 2\beta_f KN^{-\gamma+\delta} \qquad \text{for all } i/N \in S^{(2)}, i/n > \tau \\ |f(i/n) - f(\tau-)| &\leq 2\beta_f KN^{-\gamma+\delta} \qquad \text{for all } i/n \in S^{(2)}, i/n \leq \tau \end{aligned}$$

Combine this with the fact that $|\alpha - f(\tau-)|$ and $|\beta - f(\tau+)|$ are $o(1)$, which implies for sufficiently large $N$, and for any $i/n \in S^{(2)}(t)$,

$$\begin{aligned} f(i/n) - \frac{\alpha+\beta}{2} &> \frac{\Delta}{4} \qquad \text{if } i/n > \tau \\ f(i/n) - \frac{\alpha+\beta}{2} &< -\frac{\Delta}{4} \qquad \text{if } i/n \leq \tau \end{aligned}$$

The preceding fact implies that every term in the summand of (A.22) is positive, and therefore the minimizing $d$ for (A.22) would be either $\tau^{(2)} \pm rN^{-\gamma+\delta}$:

$$\begin{aligned} a(r) &= \left( \frac{\sum_{i:i/n \in S^{(2)}(t)} \left( f(i/n) - \frac{\alpha+\beta}{2} \right) \left( 1(i/n \leq \tau^{(2)} + rN^{-\gamma+\delta}) - 1(i/n \leq \tau) \right)}{\lambda_2[t - KN^{-\gamma+\delta}, t + KN^{-\gamma+\delta}]} \right) \wedge \\ &\quad \left( \frac{\sum_{i:i/n \in S^{(2)}(t)} \left( f(i/n) - \frac{\alpha+\beta}{2} \right) \left( 1(i/n \leq \tau^{(2)} - rN^{-\gamma+\delta}) - 1(i/n \leq \tau) \right)}{\lambda_2[t - KN^{-\gamma+\delta}, t + KN^{-\gamma+\delta}]} \right) \\ &\geq \frac{\Delta}{4} \cdot \frac{\lambda_2(\tau^{(2)}, \tau^{(2)} + rN^{-\gamma+\delta}] \wedge \lambda_2(\tau^{(2)} - rN^{-\gamma+\delta}, \tau^{(2)}]}{\lambda_2[t - KN^{-\gamma+\delta}, t + KN^{-\gamma+\delta}]} \end{aligned} \tag{A.24}$$

It can also be shown that for $N$ large enough (specifically $\lfloor N^{1-\gamma} \rfloor \geq 2$) and any

$d_1, d_2 \in [t - KN^{-\gamma+\delta}, t + KN^{-\gamma+\delta}]$ such that $d_2 - d_1 \geq 8/N$, we have

$$\lambda_2(d_1, d_2] \geq [N(d_2 - d_1) - 2] - \left[\frac{N(d_2 - d_1)}{\lfloor N^{1-\gamma} \rfloor} + 1\right] \geq \frac{N(d_2 - d_1)}{8}$$

In a slightly similar fashion, it can be argued that for all large $N$, $\lambda_2[t - KN^{-\gamma+\delta}, t + KN^{-\gamma+\delta}] \leq 3KN^{1-\gamma+\delta}$. Since we restricted $r$ to be greater than $8N^{-1+\gamma-\delta}$, this means

$$\begin{aligned} a(r) &\geq \frac{\Delta}{4} \cdot \frac{NrN^{-\gamma+\delta}}{8 \cdot 3KN^{1-\gamma+\delta}} \\ &\geq \frac{\Delta}{96K} r \end{aligned} \tag{A.25}$$

Hence, this shows that $a(r)$ is greater than some linear function with 0 intercept.

Now define $b(r) = (a(r) - M_n(\tau_N^{(2)}))/3 = a(r)/3$, then we have the following relation:

$$\sup_{d \in S^{(2)}} |\mathbb{M}_n(d) - M_n(d)| \leq b(r) \quad \Rightarrow \quad |\hat{\tau}_N^{(2)} - \tau_N^{(2)}| \leq rN^{-\gamma+\delta} \tag{A.26}$$

To show the above is true, suppose $d \in [t - KN^{-\gamma+\delta}, t + KN^{-\gamma+\delta}]$ and $|d - \tau_N^{(2)}| > rN^{-\gamma+\delta}$. If, in addition, the left expression above holds, then

$$\begin{aligned} \mathbb{M}_n(d) \geq M_n(d) - b(r) \geq a(r) - b(r) \quad \Rightarrow \\ \mathbb{M}_n(d) - \mathbb{M}_n(\tau^{(2)}) \geq a(r) - b(r) - M_n(\tau^{(2)}) - b(r) = b(r) > 0 \end{aligned} \tag{A.27}$$

Since $\mathbb{M}_n(d) > \mathbb{M}_n(\tau_N^{(2)})$ and $\hat{\tau}_N^{(2)}$ minimizes $\mathbb{M}_n$ among all points in $S^{(2)}(t)$, this implies $d$ could not equal $\tau_N^{(2)}$, showing that $|\hat{\tau}_N^{(2)} - \tau_N^{(2)}| \leq rN^{-\gamma}$.

Next, we bound $\mathbb{P}_{\alpha,\beta,t}\left[|\hat{\tau}_N^{(2)} - \tau_N^{(2)}| \leq rN^{-\gamma+\delta}\right]$. First, we split it into the two

parts:

$$\begin{aligned}
&\mathbb{P}_{\alpha,\beta,t}\left[|\hat{\tau}_N^{(2)} - \tau_N^{(2)}| > rN^{-\gamma+\delta}\right] \leq \\
&\mathbb{P}_{\alpha,\beta,t}\left[rN^{-\gamma+\delta} < |\hat{\tau}_N^{(2)} - \tau_N^{(2)}| \leq \eta N^{-\gamma+\delta}\right] + \mathbb{P}_{\alpha,\beta,t}\left[|\hat{\tau}_N^{(2)} - \tau_N^{(2)}| > \eta N^{-\gamma+\delta}\right] := \\
&P_N(\alpha,\beta,t) + Q_N(\alpha,\beta,t) \qquad \text{(A.28)}
\end{aligned}$$

where $\eta = K/3$. We first consider the term $P_n(\alpha,\beta,t)$. Because

$$\begin{aligned}
rN^{-\gamma+\delta} < |\hat{\tau}_N^{(2)} - \tau_N^{(2)}| \leq \eta N^{-\gamma+\delta} \quad \Rightarrow \quad & \inf_{\tau_N^{(2)}+rN^{-\gamma+\delta}<d\leq\tau_N^{(2)}+\eta N^{-\gamma+\delta}} \mathbb{M}_n(d) \leq \mathbb{M}_n(\tau) \\
\text{or} \quad & \inf_{\tau_N^{(2)}-\eta N^{-\gamma+\delta}<d\leq\tau_N^{(2)}-rN^{-\gamma+\delta}} \mathbb{M}_n(d) \leq \mathbb{M}_n(\tau), \qquad \text{(A.29)}
\end{aligned}$$

we can first split $P_n(\alpha,\beta,t)$ into the two terms

$$\begin{aligned}
P_N(\alpha,\beta,t) \quad &\leq \quad P_{N,1}(\alpha,\beta,t) + P_{N,2}(\alpha,\beta,t) \\
&=: \quad \mathbb{P}_{\alpha,\beta,t}\left[\sup_{\tau_N^{(2)}+rN^{-\gamma+\delta}<d\leq\tau_N^{(2)}+\eta N^{-\gamma+\delta}} (\mathbb{M}_n(\tau) - \mathbb{M}_n(d)) \geq 0\right] + \\
&\qquad \mathbb{P}_{\alpha,\beta,t}\left[\sup_{\tau_N^{(2)}-\eta N^{-\gamma+\delta}<d\leq\tau_N^{(2)}-rN^{-\gamma+\delta}} (\mathbb{M}_n(\tau) - \mathbb{M}_n(d)) \geq 0\right] \qquad \text{(A.30)}
\end{aligned}$$

We first form an upper bound for $P_{n,1}(\alpha,\beta,t)$ for all $(\alpha,\beta,t) \in R_n$. Note that

$$\begin{aligned}
&\mathbb{M}_n(\tau_N^{(2)}) - \mathbb{M}_n(d) \\
= \quad & -(\mathbb{M}_n(d) - M_n(d)) - M_n(d) \\
= \quad & -\frac{\sum_{i:\, i/N\in S^{(2)}(t)} \left[\left(Y_i - \frac{\alpha+\beta}{2}\right) - \left(f(i/N) - \frac{\alpha+\beta}{2}\right)\right](1(i/N \leq d) - 1(i/N \leq \tau))}{\lambda_2[t - KN^{-\gamma+\delta}, t + kN^{-\gamma+\delta}]} - M_n(d) \\
= \quad & -\frac{\sum_{i:\, i/N\in S^{(2)}(t)\cap(\tau^{(2)},d]} \varepsilon_i}{\lambda_2[t - KN^{-\gamma+\delta}, t + kN^{-\gamma+\delta}]} - \frac{\sum_{i:\, i/N\in S^{(2)}(t)\cap(\tau^{(2)},d]} \left(f(i/N) - \frac{\alpha+\beta}{2}\right)}{\lambda_2[t - KN^{-\gamma+\delta}, t + kN^{-\gamma+\delta}]} \qquad \text{(A.31)}
\end{aligned}$$

As previously explained, the $\left(f(i/N) - \frac{\alpha+\beta}{2}\right)$ term in the second summand can be bounded below by $\Delta/4$ for all sufficiently large $N$, and hence this leads to:

$$\begin{aligned} \mathbb{M}_n(\tau) - \mathbb{M}_n(d) \geq 0 \quad &\Rightarrow \\ -\sum_{i:\, i/N \in S^{(2)} \cap (\tau^{(2)}, d]} \varepsilon_i \geq \frac{\Delta}{4} &\lambda_2(\tau^{(2)}, d] \end{aligned} \tag{A.32}$$

It thus follows that

$$P_{N,1}(\alpha, \beta, t) \leq \mathbb{P}_{\alpha,\beta,t}\left[\sup_{\substack{\tau_N^{(2)} + rN^{-\gamma+\delta} < d \\ \leq \tau_N^{(2)} + \eta N^{-\gamma+\delta}}} \left(\frac{1}{\lambda_2(\tau^{(2)}, d]}\right) \left| \sum_{i:\, i/N \in S^{(2)}(t) \cap (\tau^{(2)}, d]} \varepsilon_i \right| \geq \frac{\Delta}{4}\right] \tag{A.33}$$

and by the Hajek-Renyi inequality, we get

$$\begin{aligned} &\mathbb{P}_{\alpha,\beta,t}\left[\sup_{\substack{\tau_N^{(2)} + rN^{-\gamma+\delta} < d \\ \leq \tau_N^{(2)} + \eta N^{-\gamma+\delta}}} \left(\frac{1}{\lambda_2(\tau, d]}\right) \left| \sum_{i:\, i/N \in S^{(2)}(t) \cap (\tau^{(2)}, d]} \varepsilon_i \right| \geq \frac{\Delta}{4}\right] \\ \leq\ & \frac{16}{\Delta^2}\left(\frac{1}{\lambda_2(\tau^{(2)}, \tau^{(2)} + rN^{-\gamma+\delta}]} + \sum_{j=\lambda_2(\tau^{(2)}, \tau^{(2)} + rN^{-\gamma+\delta}]}^{\lambda_2(\tau^{(2)}, \tau^{(2)} + \eta N^{-\gamma+\delta}]} \frac{1}{j^2}\right) \\ \leq\ & \frac{32}{\Delta^2} \cdot \frac{1}{\lambda_2(\tau^{(2)}, \tau^{(2)} + rN^{-\gamma+\delta}]} \end{aligned} \tag{A.34}$$

We argued earlier that $\lambda_2(\tau^{(2)}, \tau^{(2)} + rN^{-\gamma+\delta}] \geq rN^{1-\gamma+\delta}/8$ for $N$ sufficiently large enough, thus

$$P_{N,1}(\alpha, \beta, t) \leq \frac{8B}{rN^{1-\gamma+\delta}} \tag{A.35}$$

where $B = 32/\Delta^2$. From this expression we arrive at $P_{N,1}(\alpha, \beta, t) \leq \epsilon$ (for any $\epsilon > 0$) eventually, by setting $r = CN^{-1+\gamma-\delta}$ where $C$ is any constant satisfying $C > 8$ and $8B/C \leq \epsilon$.

To bound $Q_N(\alpha, \beta, t)$, from (A.25) we've argued that $a(r)$ is eventually greater than

a multiple of $r$ when $r > 8N^{-1+\gamma-\delta}$. Since we've defined $b(r) = a(r)/3$, we can find some positive constant $B'$ where $b(r) \geq B'r$ when $r > 8N^{-1+\gamma-\delta}$ (and for all large $N$). Since $\eta = K/3 > 8N^{-1+\gamma-\delta}$, eventually, this leads to

$$\begin{aligned}
& \mathbb{P}_{\alpha,\beta,t}\left[|\hat{\tau}^{(2)} - d| > \eta N^{-\gamma+\delta}\right] \\
\leq\ & \mathbb{P}_{\alpha,\beta,t}\left[\sup_{d\in S^{(2)}(t)} |\mathbb{M}_n(d) - M_n(d)| > b(\eta)\right] \\
\leq\ & \mathbb{P}_{\alpha,\beta,t}\left[\sup_{d\in S^{(2)}(t)} |\mathbb{M}_n(d) - M_n(d)| > B'\eta\right] \\
=\ & \mathbb{P}_{\alpha,\beta,t}\left[\sup_{d\in S^{(2)}(t)} \frac{\left|\sum_{i:\, i/N\in S^{(2)}(t)} \epsilon_i(\mathbb{1}(i/N \leq d) - \mathbb{1}(i/N \leq \tau))\right|}{\lambda_2(S^{(2)}(t))} > B'\eta\right] \quad \text{(A.36)}
\end{aligned}$$

Using Corollary 8.8 from *Geer* (2000), the latter expression is bounded by $C_1 \exp(-C_2\eta^2\lambda_2(S^{(2)}(t)))$ for some positive constants $C_1, C_2$, which converges to 0.

□

### A.1.4 Proof of Theorem A.2

*Proof.* Let $\{x_1^{(2)}, x_2^{(2)}, ...\}$ be the x-coordinates of the data, not used in the first stage, with corresponding response variable $(Y_1^{(2)}, Y_2^{(2)}, ...)$ and error terms $(\varepsilon_1^{(2)}, \varepsilon_2^{(2)}, ...)$. As a set, $\{x_1^{(2)}, x_2^{(2)}, ...\}$ equals $\{x_1, ..., x_N\} - \left\{\frac{\lfloor N/N_1 \rfloor}{N}, \frac{2\lfloor N/N_1 \rfloor}{N}, ...\right\}$. Note that we do not have $x_j^{(2)} = j/N$ for every integer $j$, and additionally we can write $\tau^{(2)} = x_m^{(2)}$ for some integer $m$. Since our estimate will also be one of the $x_i^{(2)}$'s, we can then denote $\hat{m}$ be the integer such that $\hat{\tau}^{(2)} = x_{\hat{m}}^{(2)}$. Note that we have the following relation between $\hat{m} - m$ and the $\lambda_2$ function on intervals:

$$\hat{m} - m = \begin{cases} \lambda_2(\tau^{(2)}, \hat{\tau}^{(2)}] & \text{when } \hat{\tau}^{(2)} > \tau \\ -\lambda_2(\hat{\tau}^{(2)}, \tau^{(2)}] & \text{when } \hat{\tau}^{(2)} \leq \tau \end{cases} \quad \text{(A.37)}$$

Hence we can write results on $\lambda_2(\tau^{(2)}, \hat{\tau}^{(2)})$ in terms of $\hat{m} - m$.

After taking a subset $S^{(2)}$ of $\{x_1^{(2)}, x_2^{(2)}, ..., x_{N-N_1}^{(2)}\}$ (specifically $S^{(2)}$ are those within $KN^{-\gamma+\delta}$ of the pilot estimate $\hat{\tau}^{(1)}$), we minimize

$$\begin{aligned}\hat{\Delta}^{(2)}(t) &:= \sum_{i:x_i \in S^{(2)}} \left( Y_i - \frac{\hat{\alpha}_N^{(1)} + \hat{\beta}_N^{(1)}}{2} \right) (1(x_i \le t) - 1(x_i \le \tau)) \\ &= \sum_{i:x_i^{(2)} \in S^{(2)}} \left( Y_i^{(2)} - \frac{\hat{\alpha}_N^{(1)} + \hat{\beta}_N^{(1)}}{2} \right) (1(x_i^{(2)} \le t) - 1(x_i^{(2)} \le \tau^{(2)})) \quad \text{(A.38)}\end{aligned}$$

over all points $t \in S^{(2)}$ to obtain the estimate for the change point. Equivalently the domain of $\hat{\Delta}^{(2)}(t)$ can be extended to all $t \in \{x_1^{(2)}, x_2^{(2)}, ...\}$, letting

$$\hat{\Delta}^{(2)}(t) = \max \left\{ \hat{\Delta}^{(2)}(r) : r \in S^{(2)} \right\} + 1 \qquad \text{for } t \notin S^{(2)}$$

The argmin of this extension is the argmin of the function restricted to $S^{(2)}$. This extended definition will be used for the next result:

**Lemma 3.** *For any fixed positive integer $j_0 > 0$,*

$$\begin{aligned}\hat{\Delta}^{(2)}\left(x_{m+j}^{(2)}\right) &= \frac{j\Delta}{2} + \epsilon_{m+1}^{(2)} + ... + \epsilon_{m+j}^{(2)} + o_p(1) \qquad \textit{for } 1 \le j \le j_0 \\ \hat{\Delta}^{(2)}\left(x_m^{(2)}\right) &= 0 + o_p(1) \\ \hat{\Delta}^{(2)}\left(x_{m-j}^{(2)}\right) &= \frac{j\Delta}{2} - \epsilon_m^{(2)} - ... - \epsilon_{m-j+1}^{(2)} + o_p(1) \qquad \textit{for } 1 \le j \le j_0 \quad \text{(A.39)}\end{aligned}$$

From this lemma it is straightforward to show the asymptotic distribution of

$\lambda_2\left(\tau_N, \hat{\tau}_N^{(2)}\right)$ is the distribution of $L_{\Delta/\sigma}$, the argmax of the random process

$$X_{\Delta/\sigma}(j) = \begin{cases} \frac{|j|\Delta}{2} - \varepsilon_{-1}^* - ... - \varepsilon_j^* & \text{, for } j < 0 \\ 0 & \text{, for } j = 0 \\ \frac{j\Delta}{2} + \varepsilon_1^* + ... + \varepsilon_j^* & \text{, for } j > 0 \end{cases} \tag{A.40}$$

where the $\{\varepsilon_j\}_{j\in\mathbb{Z}}$ are iid $N(0,\sigma^2)$ random variables.

For any fixed $\epsilon > 0$ and integer $j$, we will show that $\left|\mathbb{P}\left[\hat{m} - m = j\right] - \mathbb{P}\left[L_{\Delta/\sigma} = j\right]\right|$ $\leq \epsilon$ for all sufficiently large $N$. To do this we will first establish 3 probability bounds.

**First Bound**: First we will show that with high probability we can approximate the stochastic process $L_{\Delta/\sigma}$, which has support $\mathbb{Z}$, with a stochastic process $L_{\Delta/\sigma}^{(k)}$, which has a finite support $\mathbb{Z} \cap [-k, k]$.

We note that there exists an integer $j_1 > |j|$, such that $|L_{\Delta/\sigma}| > j_1$ with probability less than $\epsilon/3$. For any integer $k$ with $k \geq j_1$, define $L_{\Delta/\sigma}^{(k)} := \underset{|i|\leq k}{\arg\min}\{X_{\Delta/\sigma}(i)\}$. In the case that $|L_{\Delta/\sigma}| \leq k$, we have $L_{\Delta/\sigma}^{(k)} = L_{\Delta/\sigma}$, and using this we can show that $\mathbb{P}[L_{\Delta/\sigma} = j]$ is within $\epsilon/3$ of $\mathbb{P}[L_{\Delta/\sigma}^{(k)} = j]$:

$$\begin{aligned} & \left|\mathbb{P}[L_{\Delta/\sigma} = j] - \mathbb{P}[L_{\Delta/\sigma}^{(k)} = j]\right| \\ = & \left|\mathbb{P}\left[L_{\Delta/\sigma} = j, |L_{\Delta/\sigma}| \leq k\right] - \mathbb{P}\left[L_{\Delta/\sigma}^{(k)} = j, |L_{\Delta/\sigma}| \leq k\right] - \mathbb{P}\left[L_{\Delta/\sigma}^{(k)} = j, |L_{\Delta/\sigma}| > k\right]\right| \\ \leq & \left|\mathbb{P}\left[L_{\Delta/\sigma} = j, |L_{\Delta/\sigma}| \leq k\right] - \mathbb{P}\left[L_{\Delta/\sigma}^{(k)} = j, |L_{\Delta/\sigma}| \leq k\right]\right| + \mathbb{P}[|L_{\Delta/\sigma}| > k] \\ = & \left|\mathbb{P}\left[L_{\Delta/\sigma}^{(k)} = j, |L_{\Delta/\sigma}| \leq k\right] - \mathbb{P}\left[L_{\Delta/\sigma}^{(k)} = j, |L_{\Delta/\sigma}| \leq k\right]\right| + \mathbb{P}[|L_{\Delta/\sigma}| > k] \\ \leq & \ 0 + \frac{\epsilon}{3} \end{aligned} \tag{A.41}$$

**Second Bound**: We will show that there exists an integer $j_0 > j_1$ such that $|\hat{m} - m| \leq j_0$ with probability greater than $1 - \frac{\epsilon}{3}$. From our theorem on the rate of

convergence, we can find some integer $j_0 > j_1$ such that for all sufficiently large $N$,

$$\mathbb{P}\left[|\hat{\tau}^{(2)} - \tau| \leq \frac{j_0 - 2}{N}\right] > 1 - \frac{\epsilon}{3}. \tag{A.42}$$

When $|\hat{\tau}^{(2)} - \tau| \leq \frac{j_0-2}{N}$, we have $|\hat{m} - m| \leq j_0$; first we can show

$$\begin{aligned} \left|x^{(2)}_{\hat{m}} - x^{(2)}_{m}\right| &\leq \left|\hat{\tau}^{(2)} - \tau\right| + \left|\tau - \tau^{(2)}\right| \\ &\leq \frac{j_0 - 2}{N} + \frac{2}{N} \\ &= \frac{j_0}{N}, \end{aligned} \tag{A.43}$$

and second, because the $\left\{x^{(2)}_1, x^{(2)}_2, ...\right\}$ grid is just the equally spaced $\{1/N, 2/N, \cdots, N/N\}$ with some points taken out, the result of (A.43) implies $|\hat{m} - m| \leq j_0$. Hence

$$\begin{aligned} \mathbb{P}\left[|\hat{m} - m| \leq j_0\right] &\geq \mathbb{P}\left[|\hat{\tau}^{(2)} - \tau| \leq \frac{j_0 - 2}{N}\right] \\ &> 1 - \frac{\epsilon}{3} \end{aligned} \tag{A.44}$$

**Third Inequality**: Define $\hat{\tau}^{(2)}_{j_0}$ to be the minimizer of $\hat{\Delta}^{(2)}(\cdot)$ on the set $\left\{x^{(2)}_{m-j_0}, x^{(2)}_{m-j_0+1}, ..., x^{(2)}_{m+j_0}\right\}$, and let $\hat{m}_{j_0}$ be its corresponding index such that $\hat{\tau}^{(2)}_{j_0} = x^{(2)}_{\hat{m}_{j_0}}$. In the case when $|\hat{m} - m| \leq j_0$, then $\hat{\tau}^{(2)}_{j_0}$ would be equal to $\hat{\tau}^{(2)}$, and $\hat{m} = \hat{m}_{j_0}$. Using this notation we can obtain the following bound:

$$\begin{aligned} &\left|\mathbb{P}\left[\hat{m} - m = j\right] - \mathbb{P}\left[\hat{m}_{j_0} - m = j\right]\right| \\ &= \left|\mathbb{P}\left[\hat{m} - m = j, |\hat{m} - m| \leq j_0\right]\right. \\ &\left. - \mathbb{P}\left[\hat{m}_{j_0} - m = j, |\hat{m} - m| \leq j_0\right] - \mathbb{P}\left[\hat{m}_{j_0} - m = j, |\hat{m} - m| > j_0\right]\right| \\ &= \left|\mathbb{P}\left[\hat{m}_{j_0} - m = j, |\hat{m} - m| \leq j_0\right]\right. \end{aligned}$$

$$
\begin{aligned}
&- \mathbb{P}\left[\hat{m}_{j_0} - m = j, |\hat{m} - m| \leq j_0\right] - \mathbb{P}\left[\hat{m}_{j_0} - m = j, |\hat{m} - m| > j_0\right]\Bigg| \\
&\leq \mathbb{P}\left[|\hat{m} - m| > j_0\right] \\
&\leq \epsilon/3
\end{aligned}
\tag{A.45}
$$

Consider the stochastic process $\hat{\Delta}^{(2)}(x^{(2)}_{m+i})$ for $i \in \{-j_0, ..., 0, ..., j_0\}$. The previous lemma showed that, as a random variable in $\mathbb{R}^{2j_0+1}$, $\left(\hat{\Delta}^{(2)}(x^{(2)}_{m-j_0}), ..., \hat{\Delta}^{(2)}(x^{(2)}_{m+j_0})\right)$ converges in distribution to

$$(X_{\Delta/\sigma}(-j_0), ..., X_{\Delta/\sigma}(j_0)).$$

Also consider the function $\text{Ind}_{min} : \mathbb{R}^{2j_0+1} \to \mathbb{Z}$, defined as

$$\text{Ind}_{min}(a_1, ..., a_{2j_0+1}) = \left(\underset{i=1,...,2j_0+1}{\arg\min}\ (a_i)\right) - (j_0 + 1). \tag{A.46}$$

It can be easily checked that $\text{Ind}_{min}$ is a continuous function, and by definition, we also have

$$
\begin{aligned}
L^{j_0}_{\Delta/\sigma} &= \text{Ind}_{min}(X_{\Delta/\sigma}(-j_0), ..., X_{\Delta/\sigma}(j_0)) \\
\hat{m}_{j_0} - m &= \text{Ind}_{min}\left(\hat{\Delta}^{(2)}(x^{(2)}_{m-j_0}), ..., \hat{\Delta}^{(2)}(x^{(2)}_{m+j_0})\right).
\end{aligned}
\tag{A.47}
$$

Hence, by the continuous mapping theorem we have $\hat{m}_{j_0} - m$ converging to $L^{j_0}_{\Delta/\sigma}$ in distribution. For sufficiently large $N$, the absolute difference between $\mathbb{P}[L^{j_0}_{\Delta/\sigma} = j]$ and $\mathbb{P}[\hat{m}_{j_0} - m = j]$ will be less than $\epsilon/3$.

Combining what we have just shown, for sufficiently large $N$ we have

$$\begin{aligned}
& \left|\mathbb{P}\left[\hat{m}-m=j\right]-\mathbb{P}\left[L_{\Delta/\sigma}=j\right]\right| \\
\leq\ & \left|\mathbb{P}\left[\hat{m}-m=j\right]-\mathbb{P}\left[\hat{m}_{j_0}-m=j\right]\right|+\left|\mathbb{P}\left[\hat{m}_{j_0}-m=j\right]-\mathbb{P}\left[L^{j_0}_{\Delta/\sigma}=j\right]\right| \\
& +\left|\mathbb{P}\left[L^{j_0}_{\Delta/\sigma}=j\right]-\mathbb{P}[L_{\Delta/\sigma}=j]\right| \\
\leq\ & \frac{\epsilon}{3}+\frac{\epsilon}{3}+\frac{\epsilon}{3} \qquad \text{(A.48)}
\end{aligned}$$

□

**Proof of Lemma 3**

*Proof.* First note that with probability increasing to 1, $x^{(2)}_{m-j_0}, x^{(2)}_{m-j_0+1}, ..., x^{(2)}_{m+j_0}$ are all contained inside $S^{(2)}$, and this fact will be shown first. Since $\hat{\tau}^{(1)}-\tau=O_p(N^{-\gamma})$, for any $\epsilon>0$ it is possible to find a constant $C>0$ such that

$$\mathbb{P}\left[\hat{\tau}^{(1)}-CN^{-\gamma}\leq\tau\leq\hat{\tau}^{(1)}+CN^{-\gamma}\right]>1-\epsilon \qquad \text{(A.49)}$$

for all sufficiently large $N$. Additionally, for all sufficiently large $N$ we have $\frac{4+j_0}{N}\leq(KN^{\delta}-C)N^{-\gamma}$, and which means that if $|\tau^{(1)}-\tau|\leq CN^{-\gamma}$ then

$$\begin{aligned}
\hat{\tau}^{(1)}-KN^{-\gamma+\delta} &= \hat{\tau}^{(1)}-(KN^{-\gamma}-C)N^{-\gamma}-CN^{-\gamma} \\
&\leq \tau-\frac{4+2j_0}{N} \\
\hat{\tau}^{(1)}+KN^{-\gamma+\delta} &= \hat{\tau}^{(1)}+(KN^{-\gamma}-C)N^{-\gamma}+CN^{-\gamma} \\
&\geq \tau+\frac{4+2j_0}{N} \qquad \text{(A.50)}
\end{aligned}$$

Finally, for all sufficiently large $N$, we have $\lfloor N^{1-\gamma}\rfloor>2$, i.e. the first stage subsample

chooses points which are spaced more than $2/N$ points apart. Hence,

$$\begin{aligned} x^{(2)}_{m-j_0} &\geq x^{(2)}_m - 2\left(\frac{j_0+2}{N}\right) = \tau^{(2)} - 2\left(\frac{j_0+2}{N}\right) \\ &\geq \tau - \frac{2j_0+4}{N} \\ x^{(2)}_{m+j_0} &\leq x^{(2)}_m + 2\left(\frac{j_0+2}{N}\right) \\ &\leq \tau + \frac{2j_0+4}{N}, \end{aligned} \tag{A.51}$$

which leads to the conclusion that for all $N$ large enough, we have

$$\begin{aligned} 1-\epsilon &< \mathbb{P}\left[\hat{\tau}^{(1)} - CN^{-\gamma} \leq \tau \leq \hat{\tau}^{(1)} + CN^{-\gamma}\right] \\ &\leq \mathbb{P}\left[\hat{\tau}^{(1)} - KN^{-\gamma+\delta} \leq \tau - \frac{2j_0+4}{N} < \tau + \frac{2j_0+4}{N} \leq \hat{\tau}^{(1)} + KN^{-\gamma+\delta}\right] \\ &\leq \mathbb{P}\left[x^{(2)}_{m-j_0} \geq \hat{\tau}^{(1)} - KN^{-\gamma+\delta} \text{ and } x^{(2)}_{m+j_0} \leq \hat{\tau}^{(1)} + KN^{-\gamma+\delta}\right] \\ &= \mathbb{P}\left[x^{(2)}_{m-j_0} \text{ and } x^{(2)}_{m+j_0} \text{ are in } S^{(2)}\right] \end{aligned} \tag{A.52}$$

Therefore, consider the case for which $x^{(2)}_{m-j_0}$ through $x^{(2)}_{m+j_0}$ are contained in $S^{(2)}$. Under this condition we have $\hat{\Delta}^{(2)}\left(x^{(2)}_m\right) = 0$ by simple calculation, and for any $0 < j \leq j_0$,

$$\begin{aligned} & \left|\hat{\Delta}^{(2)}(x^{(2)}_{m+j}) - \left(\frac{j\Delta}{2} + \sum_{i=1}^{j}\epsilon^{(2)}_{m+i}\right)\right| \\ = & \left|\sum_{i:x^{(2)}_i \in S^{(2)} \cap \left(x^{(2)}_m, x^{(2)}_{m+j}\right]} \left(Y^{(2)}_i - \frac{\hat{\alpha}_N + \hat{\beta}_N}{2}\right) - \left(\frac{j\Delta}{2} + \sum_{i=1}^{j}\epsilon^{(2)}_{m+i}\right)\right| \\ = & \left|\sum_{i=1}^{j}\left(f(x^{(2)}_{m+i}) - \frac{\hat{\alpha}^{(1)}_N + \hat{\beta}^{(1)}_N}{2}\right) - \frac{j\Delta}{2}\right| \end{aligned}$$

$$
\begin{aligned}
&= \left| \sum_{i=1}^{j} \left[ \left( f(x_{m+i}^{(2)}) - f(\tau+) \right) + \left( \frac{f(\tau+) + f(\tau-)}{2} - \frac{\hat{\alpha}_N^{(1)} + \hat{\beta}_N^{(1)}}{2} \right) + \right. \right. \\
&\quad \left. \left. \left( \frac{f(\tau+) - f(\tau-)}{2} - \frac{\Delta}{2} \right) \right] \right| \\
&\leq \left| \sum_{i=1}^{j} \left( f(x_{m+i}^{(2)}) - f(\tau+) \right) \right| + \frac{j_0}{2} \left( \left| \hat{\alpha}_N^{(1)} - f(\tau-) \right| + \left| \hat{\beta}_N^{(1)} - f(\tau+) \right| \right) \quad \text{(A.53)}
\end{aligned}
$$

For the first term above, using an earlier argument we make the case that for sufficiently large $N$ we have $\lfloor N/N_1 \rfloor > 2$ and $x_{m+i}^{(2)} \leq \tau + \frac{2i+4}{N}$, hence

$$
\begin{aligned}
\left| \sum_{i=1}^{j} \left( f(x_{m+i}^{(2)}) - f(\tau+) \right) \right| &\leq \sum_{i=1}^{j} \beta_f \left| x_{m+i}^{(2)} - \tau \right| \\
&\leq \beta_f \sum_{i=1}^{j} \frac{2i+4}{N} \\
&\leq \frac{\beta_f}{N} (2j_0^2 + 4j_0) \qquad \text{(A.54)}
\end{aligned}
$$

For the second term, it was shown earlier that both $\left| \hat{\alpha}_N^{(1)} - f(\tau-) \right|$ and $\left| \hat{\beta}_N^{(1)} - f(\tau+) \right|$ are $O_p\left( h \vee \sqrt{\frac{1}{N_1 h}} \right)$, and hence, so is their sum. Overall, this shows that for sufficiently large $N$, $\left| \hat{\Delta}^{(2)}(x_{m+j}^{(2)}) - \left( \frac{j\Delta}{2} + \sum_{i=1}^{j} \epsilon_{m+i}^{(2)} \right) \right|$ is (uniformly for all $0 \leq j \leq j_0$) bounded above by the random variable

$$
\frac{\beta_f}{N} (2j_0^2 + 4j_0) + \frac{j_0}{2} \left( \left| \hat{\alpha}_N^{(1)} - f(\tau-) \right| + \left| \hat{\beta}_N^{(1)} - f(\tau+) \right| \right), \qquad \text{(A.55)}
$$

which is $o_p(1)$. Similarly, again for any $0 < j \leq j_0$,

$$
\begin{aligned}
&\left| \hat{\Delta}^{(2)}(x_{m-j}^{(2)}) - \left( \frac{j\Delta}{2} - \sum_{i=1}^{j} \epsilon_{m-i+1}^{(2)} \right) \right| \\
&\leq \left| \sum_{i=0}^{j-1} \left( f(x_{m-i}^{(2)}) - f(\tau) \right) \right| + \frac{j_0}{2} \left( \left| \hat{\alpha}_N^{(1)} - f(\tau-) \right| + \left| \hat{\beta}_N^{(1)} - f(\tau+) \right| \right) \quad \text{(A.56)}
\end{aligned}
$$

which is again uniformly bounded by the expression in (A.55).

Therefore, given any $\epsilon > 0$, we have

$$\begin{aligned}
& \mathbb{P}\left[\left|\hat{\Delta}^{(2)}(x^{(2)}_{m+j}) - \left(\frac{j\Delta}{2} + \sum_{i=1}^{j}\epsilon^{(2)}_{m+i}\right)\right| \geq \epsilon\right] \\
\leq\ & \mathbb{P}[x^{(2)}_{m-j_0} \notin S^{(2)} \text{ and/or } x^{(2)}_{m+j_0} \notin S^{(2)}] + \\
& \mathbb{P}\left[x^{(2)}_{m-j_0}, x^{(2)}_{m+j_0} \in S^{(2)}, \text{ and } \left|\hat{\Delta}^{(2)}(x^{(2)}_{m+j}) - \left(\frac{j\Delta}{2} + \sum_{i=1}^{j}\epsilon^{(2)}_{m+i}\right)\right| \geq \epsilon\right] \\
\leq\ & \mathbb{P}[x^{(2)}_{m-j_0} \notin S^{(2)} \text{ and/or } x^{(2)}_{m+j_0} \notin S^{(2)}] + \\
& \mathbb{P}\left[\frac{\beta_f}{N}(2j_0^2 + 4j_0) + \frac{j_0}{2}\left(\left|\hat{\alpha}^{(1)}_N - f(\tau-)\right| + \left|\hat{\beta}^{(1)}_N - f(\tau+)\right|\right) \geq \epsilon\right] \\
\to\ & 0 + 0 \qquad \text{for all } 0 < j \leq j_0
\end{aligned} \tag{A.57}$$

and similarly,

$$\begin{aligned}
& \mathbb{P}\left[\left|\hat{\Delta}^{(2)}(x^{(2)}_{m})\right| \geq \epsilon\right] \to 0 \\
& \mathbb{P}\left[\left|\hat{\Delta}^{(2)}(x^{(2)}_{m-j}) - \left(\frac{j\Delta}{2} + \sum_{i=1}^{j}\epsilon^{(2)}_{m-i+1}\right)\right| \geq \epsilon\right] \to 0
\end{aligned} \tag{A.58}$$

□

## A.2 Analysis and Proofs for Multiple Change Point Problem

Here we will provide proofs for the results presented in Section 2.3 of the main chapter. The model setup will be the same as that section.

### A.2.1 Proof of Theorem II.3

The theorem in question contains two equalities:

$$\begin{aligned} & \mathbb{P}\left[\hat{J}=J;\ \left|\lambda_2\left(\tau_j^{(2)},\hat{\tau}_j^{(2)}\right)\right| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \text{ for } j=1,\ldots,J\right] \\ = & \left(\prod_{j=1}^{J} P_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right)+o(1), \end{aligned} \tag{A.59}$$

and the second inequality

$$\begin{aligned} & \mathbb{P}\left[\hat{J}=J;\ \left|\lambda_2\left(\tau_j^{(2)},\hat{\tau}_j^{(2)}\right)\right| \leq Q_{(|\Delta_j|+2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \text{ for } j=1,\ldots,J\right] \\ = & \left(\prod_{j=1}^{J} P_{(|\Delta_j|+2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right)+o(1), \end{aligned} \tag{A.60}$$

Both these inequalities can be shown utilizing various probability bounds regarding the argmins of random walks of the form $X_\Delta(\cdot)$, and other similar random walks. To utilize such result, the estimators $\hat{\tau}_j^{(2)}$ must be expressed as the argmin of random walks with positive linear drifts. We thus introduce some new notation and new insight to facilitate this task.

First we define the indexing function $\pi_2 = \pi_{2,N}$, which maps the set of integers $\{1,\ldots,N\} - S^{(1)}$ to the set of integers $\{1,\ldots,N-|S^{(1)}|\}$:

$$\pi_2(k) = \sum_{j=1}^{k} 1(j \notin S^{(1)}). \tag{A.61}$$

For every $N$, $\pi_2$ is a strictly increasing bijection, and it also has the property such that for any two values $i$ and $j$ in the domain of $\pi_2$, $\lambda_2(i,j) = \pi_2(j) - \pi_2(i)$. Additionally, consider the following subset of the full data:

$$Y_{\pi_2^{-1}(1)}, Y_{\pi_2^{-1}(2)}, \ldots, Y_{\pi_2^{-1}(N-|S^{(1)}|)}. \tag{A.62}$$

These all the datapoints which were not used in the first stage subsample. This subset of the full dataset is also a change point model following conditions (M1) to (M4) with the same signal upper bound $\bar{\theta}$, the signal jump lower bound $\underline{\Delta}$, and the same error distribution. The change points for this subset is $\pi_2(\tau_j^{(2)})$ for $j = 1, \ldots, J$. In fact, (A.62) and the first stage subsample $\{Y_i\}_{i \in S^{(1)}}$ can be considered statistically independent change point models, resulting in the first stage estimates obtained using $\{Y_i\}_{i \in S^{(1)}}$ being independent of (A.62). This means that conditional on the first stage estimates, the distribution of (A.62) does not change from their marginal distributions.

A property of the $\pi_2$ function that will be used is a relationship between the $\pi_2^{-1}$ function and the "normal" subtraction, namely, we want to be able to compare $\pi_2^{-1}(a, b)$ to $b - a$ for $a, b \in \{1, \ldots, N - |S^{(1)}|\}$. For all $N$ large enough such that the distance between consecutive points in $S^{(1)}$ is more than 2 (i.e. $\min_{i,j \in S^{(1)}, i \neq j} |i - j| > 2$), we have the following property: for any integers $a$ and $b$ such that $a$ and $a + b$ are in $\{1, \ldots, N - |S^{(1)}|\}$,

$$\left|\pi_2^{-1}(a + b) - \pi_2^{-1}(a)\right| \leq 2|b|. \tag{A.63}$$

One way to see this is the following: for any $c, d \in \{1, \ldots, N\} - S^{(1)}$, $|\lambda_2(c, d)|$ counts the number of points in either $(c, d]$ (if $c \leq d$) or $(d, c]$ (if otherwise), and therefore $|\lambda_2(c, d)| \geq |d - c|/2$ due to the $S^{(1)}$ containing no two points that are less than 2 apart. Knowing this, the fact that

$$\left|\lambda_2\left(\pi_2^{-1}(a), \pi_2^{-1}(a + b)\right)\right| = \left|\pi_2\left(\pi_2^{-1}(a + b)\right) - \pi_2\left(\pi_2^{-1}(a)\right)\right| = |b|, \tag{A.64}$$

means that

$$2|b| = 2\left|\lambda_2\left(\pi_2^{-1}(a), \pi_2^{-1}(a+b)\right)\right| \geq \left|\pi_2^{-1}(a+b) - \pi_2^{-1}(a)\right| \tag{A.65}$$

We will offer a proof of the inequality given in (A.59). The steps for verifying (A.60) would only require a few modifications.

*Proof.* Let $\mathcal{R}_N$ be the event

$$\mathcal{R}_N := \left\{\hat{J} = J;\ \max_{j=1,\ldots,J}\left|\hat{\tau}_j^{(1)} - \tau_j\right| \leq w(N);\ \max_{j=0,\ldots,J}\left|\hat{\nu}_j^{(1)} - \nu_j\right| \leq \rho_N\right\} \tag{A.66}$$

Again define $G_N()$ joint distribution of the first stage estimates $(\hat{J}, \hat{\boldsymbol{\tau}}^{(\mathbf{1})}, \hat{\boldsymbol{\nu}}^{(\mathbf{1})}) = (\hat{J}, \hat{\tau}_1^{(1)}, \ldots, \hat{\tau}_{\hat{J}}^{(1)}, \hat{\nu}_0^{(1)}, \ldots, \hat{\nu}_{\hat{J}}^{(1)})$.

$$\begin{aligned}
& \mathbb{P}\left[\hat{J} = J;\ \left|\lambda_2\left(\tau_j^{(2)}, \hat{\tau}_j^{(2)}\right)\right| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \text{ for } j = 1, \ldots, J\right] \\
= \ & \mathbb{P}\left[\hat{J} = J;\ \left|\lambda_2\left(\tau_j^{(2)}, \hat{\tau}_j^{(2)}\right)\right| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\, \forall j;\ \mathcal{R}_N\right] \\
& + \mathbb{P}\left[\hat{J} = J;\ \left|\lambda_2\left(\tau_j^{(2)}, \hat{\tau}_j^{(2)}\right)\right| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\, \forall j;\ \text{not } \mathcal{R}_N\right] \\
= \ & \int_{(k,\boldsymbol{t},\boldsymbol{\alpha})\in\mathcal{R}_N} \mathbb{P}\Big[\max_{j=1,\ldots,J}\left|\lambda_2\left(\tau_j^{(2)}, \hat{\tau}_j^{(2)}\right)\right| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\Big|\hat{J} = k;\ \hat{\boldsymbol{\tau}}^{(\mathbf{1})} = \boldsymbol{t}; \\
& \hat{\boldsymbol{\nu}}^{(\mathbf{1})} = \boldsymbol{v}\Big]\, dG_N(k, \boldsymbol{t}, \boldsymbol{v}) \\
& + \mathbb{P}\left[\hat{J} = J;\ \left|\lambda_2\left(\tau_j^{(2)}, \hat{\tau}_j^{(2)}\right)\right| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\, \forall j;\ \text{not } \mathcal{R}_N\right]
\end{aligned}$$

Because the probability of $\mathcal{R}_N$ goes to 1, the difference between this probability and

$\prod_{j=1}^{J} P_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})$ is

$$\int_{(k,\boldsymbol{t},\boldsymbol{\alpha})\in\mathcal{R}_N} \left( \mathbb{P}\left[ \max_{j=1,\ldots,J} \left| \lambda_2\left(\tau_j^{(2)}, \hat{\tau}_j^{(2)}\right)\right| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \right| \right.$$
$$\left. \hat{J}=k;\, \hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t};\, \hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v} \right]$$
$$\left. - \prod_{j=1}^{J} P_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \right) dG_N(k,\boldsymbol{t},\boldsymbol{v}) + o(1) \tag{A.67}$$

It is therefore sufficient to show that the difference inside the integral is, for all $(k,\boldsymbol{t},\boldsymbol{\alpha})\in\mathcal{R}_N$, uniformly bounded in absolute value by a $o(1)$ term. In other words, show there is a sequence $C_{N,\alpha}=o(1)$ such that

$$\left| \mathbb{P}\left[ \max_{j=1,\ldots,J} \left| \lambda_2\left(\tau_j^{(2)}, \hat{\tau}_j^{(2)}\right)\right| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \right| \hat{J}=k;\, \hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t};\, \hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v} \right]$$
$$\left. - \prod_{j=1}^{J} P_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \right| \leq C_{N,\alpha} \tag{A.68}$$

for all admissible $(k,\boldsymbol{t},\boldsymbol{v})\in\mathcal{R}_N$.

Henceforth, consider only such admissible $k$ (which restricts to $k=J$), $\boldsymbol{t}$'s and $\boldsymbol{v}$'s, and additionally suppose that $N$ is at least large enough so that $\rho_N \leq |\underline{\Delta}|/8$ and distance between consecutive points in $S^{(1)}$ is more than 2 (i.e. $\min_{i,j\in S^{(1)}, i\neq j} |i-j|>2$). We will proceed to show (A.68), by obtaining an upper bound for the following absolute difference, for every $j=1,\ldots,J$:

$$\left| \mathbb{P}\left[ \left| \hat{\tau}_j^{(2)} - \tau_j \right| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \right| \hat{J}=J;\, \hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t};\, \hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v} \right]$$
$$\left. - P_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \right| \tag{A.69}$$

This upper bound will be derived in several components.

**First Component:** A more explicit expression for the change point estimates is

$$\begin{aligned}\hat{\tau}_j^{(2)} &= \underset{t\in S^{(2)}(t_i)}{\arg\min}(\operatorname{sgn}(v_j - v_{j-1})) \sum_{i\in S^{(2)}(t_i)} \left(Y_i - \frac{v_j + v_{j-1}}{2}\right)(1(i \le t) - 1(i \le \tau_j^{(2)})) \\ &=: \underset{t\in S^{(2)}(t_i)}{\arg\min} \hat{X}_j^{(2)}(t) \end{aligned} \tag{A.70}$$

Since $N$ was assumed to be large enough so that $\rho_N < |\underline{\Delta}|/8 \le (\nu_j - \nu_{j-1})/8$, the sign of $v_j - v_{j-1}$ is the same as the sign of $\Delta_j := \nu_j - \nu_{j-1}$, making the optimized

expression above equal

$$\hat{X}_j^{(2)}(t) = \operatorname{sgn}(\Delta_j) \sum_{i \in S^{(2)}(t_i)} \left( Y_i - \frac{v_j + v_{j-1}}{2} \right) (1(i \le t) - 1(i \le \tau_j^{(2)}))$$

$$= \begin{cases} \left| (\tau_j^{(2)}, t] \cap S^{(2)}(t_j) \right| \left[ \frac{|\Delta_j|}{2} + \operatorname{sgn}(\Delta_j) \hat{D}_j \right] + & \\ \operatorname{sgn}(\Delta_j) \sum\limits_{i \in (\tau_j^{(2)}, t] \cap S^{(2)}(t_j)} \varepsilon_i & \text{for } t > \tau_j \\ 0 & t = \tau_j \\ \left| (t, \tau_j^{(2)}] \cap S^{(2)}(t_j) \right| \left[ \frac{|\Delta_j|}{2} - \operatorname{sgn}(\Delta_j) \hat{D}_j \right] - & \\ \operatorname{sgn}(\Delta_j) \sum\limits_{i \in (t, \tau_j^{(2)}] \cap S^{(2)}(t_j)} \varepsilon_i & \text{for } t > \tau_j \end{cases}$$

$$= \begin{cases} \left( \pi_2(t) - \pi_2(\tau_j^{(2)}) \right) \cdot \left[ \frac{|\Delta_j|}{2} + \operatorname{sgn}(\Delta_j) \hat{D}_j \right] + & \\ \operatorname{sgn}(\Delta_j) \sum\limits_{i=\pi_2(\tau_j^{(2)})+1}^{\pi_2(t)} \varepsilon_{\pi_2^{-1}(i)} & \text{for } t > \tau_j \\ 0 & t = \tau_j \\ \left( \pi_2(\tau_j^{(2)}) - \pi_2(t) \right) \cdot \left[ \frac{|\Delta_j|}{2} - \operatorname{sgn}(\Delta_j) \hat{D}_j \right] - & \\ \operatorname{sgn}(\Delta_j) \sum\limits_{i=\pi_2(t)+1}^{\pi_2(\tau_j^{(2)})} \varepsilon_{\pi_2^{-1}(i)} & \text{for } t < \tau_j \end{cases}$$

$$\text{since } (a, b] \cap S^{(2)}(t_j) = \pi_2^{-1}\left( (\pi_2(a), \pi_2(b)] \right) \text{ for any } (a, b] \subset S^{(2)}(t_j) \tag{A.71}$$

where

$$\hat{D}_j = \frac{\nu_j - v_j}{2} + \frac{\nu_{j-1} - v_{j-1}}{2} \tag{A.72}$$

which is less than $\rho_N$ in absolute value. From the equalities written above, we can

deduce that for any integer $t$ such that $\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right) \in S^{(2)}(t_j)$,

$$\begin{aligned}
&\hat{X}_j^{(2)}\left(\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\right) \\
&= \begin{cases} t\cdot\left[\frac{|\Delta_j|}{2}+\mathrm{sgn}(\Delta_j)\hat{D}_j\right]+\mathrm{sgn}(\Delta_j)\sum\limits_{i=\pi_2(\tau_j^{(2)})+1}^{\pi_2(\tau_j^{(2)})+t}\varepsilon_{\pi_2^{-1}(i)} & \text{for } t>\tau_j \\ 0 & t=\tau_j \\ |t|\cdot\left[\frac{|\Delta_j|}{2}-\mathrm{sgn}(\Delta_j)\hat{D}_j\right]-\mathrm{sgn}(\Delta_j)\sum\limits_{i=\pi_2(\tau_j^{(2)})+t+1}^{\pi_2(\tau_j^{(2)})}\varepsilon_{\pi_2^{-1}(i)} & \text{for } t<\tau_j \end{cases}
\end{aligned} \tag{A.73}$$

This allows us to compare $\hat{X}_j^{(2)}$ with the random walk $X'_{|\Delta_j|-2|\hat{D}_j|}$:

$$X'_{|\Delta_j|-2|\hat{D}_j|}(t) := \begin{cases} t\frac{|\Delta_j|-2|\hat{D}_j|}{2}+\mathrm{sgn}(\Delta_j)\sum\limits_{i=1}^{t}\varepsilon_{\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+i\right)} & \text{for } t>0 \\ 0 & t=0 \\ |t|\frac{|\Delta_j|-2|\hat{D}_j|}{2}-\mathrm{sgn}(\Delta_j)\sum\limits_{i=t}^{-1}\varepsilon_{\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+i+1\right)} & \text{for } t<0 \end{cases} \tag{A.74}$$

Specifically, the random process $\hat{X}_j^{(2)}\left(\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\right)$ either equals the sum $X'_{|\Delta_j|-2|\hat{D}_j|}(t)$ $+2|\hat{D}_j t|1(t>0)$ or $X'_{|\Delta_j|-2|\hat{D}_j|}(t)+2|\hat{D}_j t|1(t<0)$. In either case, this begs the use of Lemma 10 and Lemma 13, but first we must verify some the conditions of those results, namely that $|\Delta_j|>2|\hat{D}_j|$ (automatically true since $N$ was assumed to be large enough so that $|\Delta_j|\geq\underline{\Delta}\geq 8\rho_N$), and secondly, $\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}(t_j)$ for $|t|\leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})$. To see that this is true, first use (A.63) to arrive at

$$\begin{aligned}
&\left|\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})\pm Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right)-\tau_j^{(2)}\right| \\
=\;&\left|\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})\pm Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right)-\pi_2^{-1}\left(\pi_2\left(\tau_j^{(2)}\right)\right)\right| \\
\leq\;& 2Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})
\end{aligned} \tag{A.75}$$

Therefore, given any $|t| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})$ we have

$$\pi_2^{-1}\left(\pi_2\left(\tau_j^{(2)}\right)+t\right) \in \left[\tau_j^{(2)} \pm 2Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right] - S^{(1)}.$$

Next, due to Lemma 8, there are positive expressions $C_1(\cdot)$, $C_2(\cdot)$, both decreasing, such that

$$\begin{aligned} Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) &\leq C_1(|\Delta_j|-2\rho_N)\log\left(\frac{C_2(|\Delta_j|-2\rho_N)J}{\alpha}\right) \\ &\leq C_1(\underline{\Delta}/2)\log\left(\frac{C_2(\underline{\Delta}/2)J}{\alpha}\right) \\ &\leq C_1(\underline{\Delta}/2)\log\left(\frac{C_2(\underline{\Delta}/2)N}{\alpha}\right) \end{aligned} \tag{A.76}$$

Since $w(N)$ is greater than order of $N^{1-\gamma}$, which in turn is greater in order than $\log(N)$, we see that for all large $N$:

$$\begin{aligned} & t_j - Kw(N) \leq \tau_j^{(2)} - (K-1)w(N) + 1 < \tau_j^{(2)} - 2Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \\ \leq\ & \pi_2^{-1}\left(\pi_2(\tau_j^{(2)}) - Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right) < \pi_2^{-1}\left(\pi_2(\tau_j^{(2)}) + Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right) \\ \leq\ & \tau_j^{(2)} + 2Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \leq \tau_j^{(2)} + (K-1)w(N) - 1 \leq t_j + Kw(N) \end{aligned} \tag{A.77}$$

Therefore given any $|t| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})$,

$$\begin{aligned} \pi_2^{-1}\left(\pi_2\left(\tau_j^{(2)}\right)+t\right) &\in \left[\tau_j^{(2)} \pm 2Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right] - S^{(1)} \\ &\subset [t_j \pm Kw(N)] - S^{(1)} \\ &= S^{(2)}(t_j) \end{aligned} \tag{A.78}$$

showing that the conditions of Lemma 10 are satisfied. Before continuing, we make a small point before continuing. Note that for any integer $t^* \in S^{(2)}(t_j)$, $\lambda_2\left(\tau_j^{(2)}, \hat{\tau}_j^{(2)}\right) = t^*$ is an equivalent event to $\underset{t:\, \pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}}{\arg\min} \hat{X}_j^{(2)}\left(\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\right) = t^*$. This is

because

$$\begin{aligned}
&\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}}{\arg\min}\ \hat{X}_j^{(2)}\left(\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\right)=t^* \\
\longleftrightarrow\ &\underset{t\in S^{(2)}(t_j)}{\arg\min}\hat{X}_j^{(2)}(t)=\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t^*\right) \\
\longleftrightarrow\ &\pi_2\left(\underset{t\in S^{(2)}(t_j)}{\arg\min}\hat{X}_j^{(2)}(t)\right)-\pi_2(\tau_j^{(2)})=t^* \\
\longleftrightarrow\ &\lambda_2\left(\tau_j^{(2)},\hat{\tau}_j^{(2)}\right)=t^* \qquad\qquad (A.79)
\end{aligned}$$

We are now ready to apply Lemma 10:

$$\begin{aligned}
&\mathbb{P}\left[\left|\lambda_2\left(\tau_j^{(2)},\hat{\tau}_j^{(2)}\right)\right|\le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\middle|\hat{J}=J,\hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t},\,\hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v}\right] \\
&-\mathbb{P}\left[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}}{\arg\min}X'_{|\Delta_j|-2|\hat{D}_j|}(t)\right|\le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right| \qquad (A.80) \\
&\hat{J}=J,\hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t},\,\hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v}\Bigg] \\
=\ &\mathbb{P}\Bigg[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}}{\arg\min}\sigma^{-1}\hat{X}_j^{(2)}\left(\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\right)\right|\le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \\
&\left|(\hat{J},\hat{\boldsymbol{\tau}}^{(\mathbf{1})},\hat{\boldsymbol{\nu}}^{(\mathbf{1})})=(J,\boldsymbol{t},\boldsymbol{v})\right.\Bigg] \\
&-\mathbb{P}\Bigg[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}}{\arg\min}\sigma^{-1}X'_{|\Delta_j|-2|\hat{D}_j|}(t)\right|\le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \\
&\left|\hat{J}=J,\hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t},\,\hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v}\right.\Bigg] \\
&\text{indexing change by (A.79); and division by constant } \sigma \text{ does not change argmin value} \\
\le\ &2\hat{D}_j\sigma^{-1}\Big[A_1^+\left((|\Delta_j|-2\hat{D}_j)/\sigma\right)\left(Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right)^{3/2} \\
&+B_1^+\left((|\Delta_j|-2\hat{D}_j)/\sigma\right)\sqrt{Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})}\Big] \\
&\times\exp\left[-C_1^+\left((|\Delta_j|-2\hat{D}_j)/\sigma\right)Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]
\end{aligned}$$

Because, as stated previously, $\hat{X}_j^{(2)}\left(\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\right)$ either equals $X'_{|\Delta_j|-2|\hat{D}_j|}(t)+2|\hat{D}_j t|1(t>0)$ or $X'_{|\Delta_j|-2|\hat{D}_j|}(t)+2|\hat{D}_j t|1(t<0)$, Lemma 10 leads to this inequality for some positive monotone expressions $A_1^+()$, $B_1^+()$, and $C_1^+()$, which are decreasing, decreasing, and increasing. This expression could be further bounded by footnotesize

$$\leq 2\sigma^{-1}\hat{D}_j\left[A_1^+\left(\underline{\Delta}/2\sigma\right)\left(Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right)^{3/2}+B_1^+\left(\underline{\Delta}/2\sigma\right)\sqrt{Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})}\right]$$
$$\times\exp\left[-C_1^+\left(\underline{\Delta}/2\sigma\right)Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]$$
$$\leq 2C^*\left(\underline{\Delta}/2\sigma\right)\rho_N$$

where $C_+^*(\cdot)=\sup_{x\in\mathbb{R}^+}\left([A(\cdot)x^{3/2}+B(\cdot)\sqrt{x}]\exp(-C(\cdot)x)\right)$, guaranteed to be finite

(A.81)

In a similar manner, apply Lemma 13 to obtain

$$\mathbb{P}\left[\left|\lambda_2\left(\tau_j^{(2)},\hat{\tau}_j^{(2)}\right)\right|\leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\middle|\hat{J}=J,\hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t},\hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v}\right]$$
$$-\mathbb{P}\left[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}}{\arg\min}X'_{|\Delta_j|-2|\hat{D}_j|}(t)\right|\leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\middle|\hat{J}=J,\hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t},\hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v}\right]$$
$$=\mathbb{P}\left[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}}{\arg\min}\frac{1}{\sigma}\hat{X}_j^{(2)}\left(\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\right)\right|\leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right.$$
$$\left.\middle|\hat{J}=J,\hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t},\hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v}\right]$$

$$
\begin{aligned}
&- \mathbb{P}\left[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}}{\arg\min}\frac{1}{\sigma}X'_{|\Delta_j|-2|\hat{D}_j|}(t)\right| \le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\middle|\hat{J}=J,\hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t},\hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v}\right]\\
&\ge -2A_1^-\left(\left(\frac{|\Delta_j|}{2}-3|\hat{D}_j|\right)/\sigma\right)\rho_N\sqrt{Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})}\\
&\times\exp\left[-B_1^-\left(\left(\frac{|\Delta_j|}{2}-3|\hat{D}_j|\right)/\sigma\right)Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]\\
&\text{for some positive decreasing expression } A_1^- \text{ and some positive increasing expression } B_1^-\\
&\ge -2A_1^-\left(\underline{\Delta}/2\sigma\right)\rho_N\sqrt{Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})}\exp\left[-B_1^-\left(\underline{\Delta}/2\sigma\right)Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]\\
&\ge -2C_1^-(\underline{\Delta}/2\sigma)\rho_N \quad \text{where } C_1^-(\cdot)=\sup_{x\in\mathbb{R}^+}A_1^-(\cdot)\sqrt{x}\exp\left[B_1^-(\cdot)x\right]
\end{aligned}
\tag{A.82}
$$

Altogether, both (A.81) and (A.82) together imply

$$
\begin{aligned}
&\left|\mathbb{P}\left[\left|\lambda_2\left(\tau_j^{(2)},\hat{\tau}_j^{(2)}\right)\right|\le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\middle|\hat{J}=J,\hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t},\hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v}\right]\right.\\
&\left.-\mathbb{P}\left[\left|\underset{t\in S^{(2)}(t_j)-\tau_j}{\arg\min}X'_{|\Delta_j|-2|\hat{D}_j|}(t)\right|\le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\middle|\hat{J}=J,\hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t},\hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v}\right]\right|\\
&\le 2\left(C_1^-(\underline{\Delta}/2\sigma)\vee C_1^+(\underline{\Delta}/2\sigma)\right)\rho_N
\end{aligned}
\tag{A.83}
$$

**Second Component:** Now $X'_{|\Delta|-2|\hat{D}_j|}(t)/\sigma$ has the same exact distribution as $X_{(|\Delta|-2|\hat{D}_j|)/\sigma}(t)$, for all integers $t$ such that $\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}(t_j)$. It was also shown in the previous section that the set $\left\{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}(t_j)\right\}$ contains the interval of integers $\left[\pm Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]$. Therefore apply Lemma 11 to first obtain

$$
\begin{aligned}
&\mathbb{P}\left[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}(t_j)}{\arg\min}X'_{|\Delta_j|-2|\hat{D}_j|}(t)\right|\le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\middle|\hat{J}=J,\hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t},\hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v}\right]\\
=\;&\mathbb{P}\left[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}(t_j)}{\arg\min}\frac{1}{\sigma}X'_{|\Delta_j|-2|\hat{D}_j|}(t)\right|\le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\middle|\hat{J}=J,\hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t},\hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v}\right]\\
=\;&\mathbb{P}\left[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}(t_j)}{\arg\min}X_{(|\Delta_j|-2|\hat{D}_j|)/\sigma}(t)\right|\le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]\\
\ge\;&\mathbb{P}\left[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}(t_j)}{\arg\min}X_{(|\Delta_j|-2\rho_N)/\sigma}(t)\right|\le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]
\end{aligned}
\tag{A.84}
$$

and apply Lemma 11 to obtain another equality in the other direction

$$\mathbb{P}\left[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}(t_j)}{\arg\min} X'_{|\Delta_j|-2|\hat{D}_j|}(t)\right| \le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\middle|\hat{J}=J,\hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t},\hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v}\right]$$

$$\begin{aligned}
&= \mathbb{P}\left[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}(t_j)}{\arg\min} X_{(|\Delta_j|-2|\hat{D}_j|)/\sigma}(t)\right| \le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]\\
&\le \mathbb{P}\left[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}(t_j)}{\arg\min} X_{(|\Delta_j|-2\rho_N)/\sigma}(t)\right| \le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right] +\\
&\quad (2\rho_N-2|\hat{D}_j|)\Big[A_2\left((|\Delta_j|-2\rho_N)/\sigma\right)\left(Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right)^{3/2}\\
&\quad +B_2\left((|\Delta_j|-2\rho_N)/\sigma\right)\sqrt{Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})}\Big]\\
&\quad \times\exp\left[-C_2\left(|\Delta_j|-2\rho_N\right)Q_{|\Delta_j|-2\rho_N}(\sqrt[J]{1-\alpha})\right]\\
&\le \mathbb{P}\left[\left|\underset{t\in S^{(2)}(t_j)-\tau_j}{\arg\min} X_{(|\Delta_j|-2\rho_N)/\sigma}(t)\right| \le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right] +\\
&\quad 2\rho_N\left[A_2\left(\underline{\Delta}/2\sigma\right)\left(Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right)^{3/2}+B_2\left(\underline{\Delta}/2\sigma\right)\sqrt{Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})}\right]\\
&\quad \times\exp\left[-C_2\left(\underline{\Delta}/2\sigma\right)Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]\\
&\quad \text{since } N \text{ is large enough such that } 2\rho_N<\underline{\Delta}/2, \text{ and by the monotonicity of } A_2, B_2, \text{ and } C_2\\
&\le \mathbb{P}\left[\left|\underset{t\in S^{(2)}(t_j)-\tau_j}{\arg\min} X_{(|\Delta_j|-2\rho_N)/\sigma}(t)\right| \le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right] + 2C_2^*\left(\frac{\underline{\Delta}}{2\sigma}\right)\rho_N\\
&\quad \text{where } C_2^*(\cdot)=\sup_{x\in\mathbb{R}^+}\Big(\left[A_2(\cdot)x^{3/2}+B_2(\cdot)\sqrt{x}\right]\exp(-C_2(\cdot)x)\Big) \qquad \text{(A.85)}
\end{aligned}$$

These two inequalities together imply a bound on the absolute difference:

$$\begin{aligned}
&\left|\mathbb{P}\left[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}(t_j)}{\arg\min} X'_{|\Delta_j|-2|\hat{D}_j|}(t)\right| \le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\middle|\hat{J}=J,\hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t},\hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v}\right]\right.\\
&\quad \left.-\mathbb{P}\left[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}(t_j)}{\arg\min} X_{(|\Delta_j|-2\rho_N)/\sigma}(t)\right| \le Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]\right|\\
&\quad \le 2C_2^*\left(\frac{\underline{\Delta}}{2\sigma}\right)\rho_N \qquad \text{(A.86)}
\end{aligned}$$

**Third Component:** We note that the set $\left\{t:\ \pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}(t_j)\right\}$ contains the set $\left[\pm\left(\frac{(K-1)w(N)}{2}-1\right)\right]$. This is because by (A.63),

$$\left|\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})\pm\left(\frac{(K-1)w(N)}{2}-1\right)\right)-\tau_j^{(2)}\right|\leq 2\left(\frac{(K-1)w(N)}{2}-1\right) \quad \text{(A.87)}$$

secondly because

$$\begin{aligned} & t_j-Kw(N)\leq \tau_j^{(2)}-(K-1)w(N)+1<\tau_j^{(2)}-2\left(\frac{(K-1)w(N)}{2}-1\right) \\ < \ & \tau_j^{(2)}+2\left(\frac{(K-1)w(N)}{2}-1\right)\leq \tau_j^{(2)}+(K-1)w(N)\leq t_j+Kw(N) \quad \text{(A.88)}\end{aligned}$$

Therefore, for any $t$ such that $|t|\leq\left(\frac{(K-1)w(N)}{2}-1\right)$, we have $\pi_2^{-1}(\pi_2(\tau_j^{(2)})+t)\in[t_j\pm Kw(N)]-S^{(1)}=S^{(2)}(t_j)$. This allows an application of Lemma 6 to obtain

$$\begin{aligned} & \left|\mathbb{P}\left[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}(t_j)}{\arg\min} X_{(|\Delta_j|-2\rho_N)/\sigma}(t)\right|\leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]\right. \\ & \left.-\mathbb{P}\left[\left|\underset{|t|\leq\frac{(K-1)W(N)}{2}-1}{\arg\min} X_{(|\Delta_j|-2\rho_N)/\sigma}(t)\right|\leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]\right| \\ \leq \ & A_3((|\Delta_j|-2\rho_N)/\sigma)\exp\left(-B_3((|\Delta_j|-2\rho_N)/\sigma)\left(\frac{(K-1)w(N)}{2}-1\right)\right) \\ & \text{for some decreasing expression } A_3() \text{ and increasing expression } B_3() \\ \leq \ & A_3(\underline{\Delta}/2\sigma)\exp\left(-B_3(\underline{\Delta}/2\sigma)\left(\frac{(K-1)w(N)}{2}-1\right)\right) \quad \text{(A.89)}\end{aligned}$$

The same application of the lemma can also yield

$$\begin{aligned}
&\left|\mathbb{P}\left[\left|\underset{|t|\leq\frac{(K-1)W(N)}{2}-1}{\arg\min} X_{(|\Delta_j|-2\rho_N)/\sigma}(t)\right| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]\right. \\
&\left.-\mathbb{P}\left[\left|\underset{t\in\mathbb{Z}}{\arg\min} X_{(|\Delta_j|-2\rho_N)/\sigma}(t)\right| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]\right| \\
\leq\ & A_3(\underline{\Delta}/2\sigma)\exp\left(-B'(\underline{\Delta}/2\sigma)\left(\frac{(K-1)w(N)}{2}-1\right)\right) \qquad \text{(A.90)}
\end{aligned}$$

Adding up these two upper bounds imply

$$\begin{aligned}
&\left|\mathbb{P}\left[\left|\underset{t:\,\pi_2^{-1}\left(\pi_2(\tau_j^{(2)})+t\right)\in S^{(2)}(t_j)}{\arg\min} X_{(|\Delta_j|-2\rho_N)/\sigma}(t)\right| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]\right. \\
&\left.-\mathbb{P}\left[\left|\underset{t\in\mathbb{Z}}{\arg\min} X_{(|\Delta_j|-2\rho_N)/\sigma}(t)\right| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]\right| \\
\leq\ & 2A_3(\underline{\Delta}/2\sigma)\exp\left(-B'(\underline{\Delta}/2\sigma)\left(\frac{(K-1)w(N)}{2}-1\right)\right) \qquad \text{(A.91)}
\end{aligned}$$

**Sum of the Components:** Adding up the differences in (A.83), (A.86), and (A.91):

$$\begin{aligned}
&\left|\mathbb{P}\left[\left|\lambda_2\left(\tau_j^{(2)},\hat{\tau}_j^{(2)}\right)\right| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\middle|\hat{J}=J,\hat{\boldsymbol{\tau}}^{(\mathbf{1})}=\boldsymbol{t},\hat{\boldsymbol{\nu}}^{(\mathbf{1})}=\boldsymbol{v}\right]\right. \\
&\left.-\mathbb{P}\left[\left|\underset{t\in\mathbb{Z}}{\arg\min} X_{(|\Delta_j|-2\rho_N)/\sigma}(t)\right| \leq Q_{(|\Delta_j|-2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right]\right| \\
\leq\ & C_4\rho_N + C_5\exp\left[-C_6(K-1)w(N)\right] \qquad \text{(A.92)}
\end{aligned}$$

for some constants $C_4$, $C_5$, and $C_6$.

Finally, in order to bound (A.68), we note that given the two real valued triangular arrays $a_{N,1},\ldots,a_{N,J}$ and $b_{N,1},\ldots,b_{N,J}$, all of which contained in the continuous interval $[0,1]$, such that $\left|\frac{a_{N,i}-b_{N,i}}{b_{N,i}}\right| \leq C_N$ for $1\leq i\leq J$, where $JC_N\to 0$ as $N\to\infty$,

then $|\prod a_{N,j} - \prod b_{N,j}| \to 0$. This is because

$$\left|\prod_{j=1}^{J} a_{N,j} - \prod_{j=1}^{J} b_{N,j}\right| = \left|\prod_{j=1}^{J} b_{N,j}\right| \left|\prod_{j=1}^{J}\left(1 - \frac{a_{N,j} - b_{N,j}}{b_{N,j}}\right) - 1\right| \tag{A.93}$$

Since $|\prod b_{N,j}| \in [0,1]$, the above converges to 0 if $\prod\left(1 - \frac{a_{N,j}-b_{N,j}}{b_{N,j}}\right) \to 1$, which is true since

$$\begin{aligned}
\prod_{j=1}^{J}\left(1 - \frac{a_{N,j} - b_{N,j}}{b_{N,j}}\right) &\geq (1 - C_N)^J \geq 1 - C_N J \to 1 \\
\prod_{j=1}^{J}\left(1 - \frac{a_{N,j} - b_{N,j}}{b_{N,j}}\right) &\leq (1 + C_N)^J \leq (e^{C_N})^J \to e^0 = 1
\end{aligned} \tag{A.94}$$

This result is useful because for all $N$ large enough so that $2Kw(N) \leq \delta_N$, the stage two sets $S^{(2)}(t_j)$'s for $j = 1, \ldots, J$ are mutually exclusive sets, and hence by conditional independence, (A.68) equals

$$\begin{aligned}
&\left|\prod_{j=1}^{J} \mathbb{P}\left[\left|\lambda_2\left(\tau_j^{(2)}, \hat{\tau}_j^{(2)}\right)\right| \leq Q_{(|\Delta_j| - 2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \middle| \hat{J} = J;\, \hat{\boldsymbol{\tau}}^{(\mathbf{1})} = \boldsymbol{t};\, \hat{\boldsymbol{\nu}}^{(\mathbf{1})} = \boldsymbol{v}\right]\right. \\
&\left. - \prod_{j=1}^{J} P_{(|\Delta_j| - 2\rho_N)/\sigma}(\sqrt[J]{1-\alpha})\right|
\end{aligned} \tag{A.95}$$

More-ever, using (A.92),

$$\begin{aligned}
& J\left|\mathbb{P}\left[\left|\lambda_2\left(\tau_j^{(2)}, \hat{\tau}_j^{(2)}\right)\right| \leq Q_{(|\Delta_j| - 2\rho_N)/\sigma}(\sqrt[J]{1-\alpha}) \middle| \hat{J} = J, \hat{\boldsymbol{\tau}}^{(\mathbf{1})} = \boldsymbol{t},\, \hat{\boldsymbol{\nu}}^{(\mathbf{1})} = \boldsymbol{v}\right]\right. \\
& \left. - \mathbb{P}_{|\Delta_j| - 2\rho_N}(\sqrt[J]{1-\alpha})\right| \left(\mathbb{P}_{|\Delta_j| - 2\rho_N}(\sqrt[J]{1-\alpha})\right)^{-1} \\
\leq\ & J\left(C_4\rho_N + C_5 \exp\left[-C_6(K-1)w(N)\right]\right)(1-\alpha)^{-1/J} \\
\leq\ & C_4(1-\alpha)^{-1} J\rho_N + C_5(1-\alpha)^{-1} N \exp\left[-C_6(K-1)w(N)\right],
\end{aligned} \tag{A.96}$$

which goes to 0 since $J\rho_N \to 0$ and $w(N) \geq CN^{1-\gamma}$ for some constant $C$. This lets us conclude that (A.95) converges to 0. □

### A.2.2 Proof of Theorem II.7

We will show that

$$\mathbb{P}\left[\hat{J} = J;\ \left|\hat{\tau}_j^{re} - \tau_j^{**}\right| \leq Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1 - \frac{\alpha_N}{J}\right)\ \forall j\right] \geq 1 + \alpha_N + o(1) \tag{A.97}$$

*Proof.* Letting $\mathcal{R}_N$ be the event that

$$\left\{\hat{J} = J;\ \max_{j=1,\ldots,J} |\hat{\tau}_j^* - \tau_j^{**}| \leq w^*(N^*);\ \max_{j=0,\ldots,J} |\hat{\nu}_j^{(1)} - \nu_j| \leq \rho_N\right\}. \tag{A.98}$$

In a similar way to the proof for Theorem II.3, to prove

$$\begin{aligned} &\mathbb{P}\left[\hat{J} = J;\ \max_{j=1,\ldots,J} |\hat{\tau}_j^{re} - \tau_j^{**}| \leq Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1 - \frac{\alpha_N}{J}\right)\ \text{for all } j = 1\ldots, J\right] \\ \geq\ & 1 - \alpha_N + o(1) \end{aligned} \tag{A.99}$$

it is sufficient to demonstrate that

$$\begin{aligned} &\mathbb{P}\left[\max_{j=1,\ldots,J} |\hat{\tau}_j^{re} - \tau_j^{**}| \leq Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1 - \frac{\alpha_N}{J}\right)\ \forall j \middle| J = \hat{J}, \hat{\tau}_j^* = t_j, \hat{\nu}_j^{(1)} = v_j\ \forall j\right] \\ =\ & 1 - P_{(|\Delta_j|+2\rho_N)/\sigma}\left(1 - \frac{\alpha_N}{J}\right) + o(1) \end{aligned} \tag{A.100}$$

for all $t_j$'s and $v_j$'s permissible within $\mathcal{R}_N$, which we will assume when we write $t_j$'s and $v_j$'s from here on.

We now try to bound the difference between

$$\mathbb{P}\left[|\hat{\tau}_j^{re} - \tau_j^{**}| \leq Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1 - \frac{\alpha_N}{J}\right) \middle| J = \hat{J}, \hat{\tau}_j^* = t_j, \hat{\nu}_j^{(1)} = v_j \text{ for all } j\right]$$

and $P_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right)$ for all $j$. Each estimator equals the argmin of a random walk:

$$\begin{aligned}\hat{\tau}_j^{re} &:= \underset{t\in[t_j\pm\hat{d}_j]}{\arg\min}(\mathrm{sgn}(v_j-v_{j-1}))\sum_{i\in[t_j\pm\hat{d}_j]}\left(Y_i-\frac{v_j+v_{j-1}}{2}\right)\left[1(i\le t)-1(i\le\tau_j^{**})\right]\\ &=: \underset{t\in[t_j\pm\hat{d}_j]}{\arg\min}\hat{X}_j(t). \qquad \text{(A.101)}\end{aligned}$$

In comparison with the random walk

$$X_j'(t) := \begin{cases} t\left(\frac{|\Delta_j|}{2}+\mathrm{sgn}(\Delta_j)\hat{D}_j\right)+\mathrm{sgn}(\Delta_j)\sum_{i=1}^{t}\varepsilon_{\tau_j^{**}+i} & t>0\\ 0 & t=0\\ |t|\left(\frac{|\Delta_j|}{2}-\mathrm{sgn}(\Delta_j)\hat{D}_j\right)-\mathrm{sgn}(\Delta_j)\sum_{i=1}^{t}\varepsilon_{\tau_j^{**}-i+1} & t<0\end{cases} \qquad \text{(A.102)}$$

where $\hat{D}_j=\frac{v_j-\nu_j+v_{j-1}-\nu_{j-1}}{2}$. We have, for all sufficiently large $N$, $\hat{X}_j(t+\tau_j^{**})$ equaling either $X_j'(t)-2|t|\hat{D}_j1(t>0)$ or $X'(t)-2|t|\hat{D}_j1(t<0)$ for all integers $t\in[t_j\pm\hat{d}_j]-\tau_j^{**}$. With regards to the set $[t_j\pm\hat{d}_j]-\tau_j^{**}$, for all large $N$ (such that $3w^*(N^*)\le\delta_{N^*}^*/2$) this set contains the interval $[-\delta_{N^*}^*/2,\delta_{N^*}^*/2]$. This can be seen by the series of inequalities

$$\begin{aligned}& t_j-\hat{d}_j\le\tau_j^{**}+w^*(N^*)-(\delta_{N^*}^*-2w^*(N^*))\le\tau_j^{**}-\delta_{N^*}^*/2\\ <\;& \tau_j^{**}-\delta_{N^*}^*/2\le\tau_j^{**}-w^*(N^*)+(\delta_{N^*}^*-2w^*(N^*))\le t_j+\hat{d}_j. \qquad \text{(A.103)}\end{aligned}$$

Furthermore, since $\log(J/\alpha_N)=o(\delta_{N^*}^*)$, we can use Lemma 5 to obtain, for all large $N$,

$$\begin{aligned}&\tau_j^{**}-\delta_{N^*}^*/2\le\tau_j^{**}-\frac{1}{B}\log\left(\frac{AJ}{\alpha_N}\right)\le\tau_j^{**}-Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right)\\ <\;& \tau_j^{**}+Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right)\le\tau_j^{**}+\frac{1}{B}\log\left(\frac{AJ}{\alpha_N}\right)\le\tau_j^{**}+\delta_{N^*}^*/2 \qquad \text{(A.104)}\end{aligned}$$

where $A$ and $B$ are some constants. Therefore, we can apply Lemma 10 to obtain the

inequality

$$\begin{aligned}
&\mathbb{P}\left[|\hat{\tau}_j^{re} - \tau_j^{**}| \leq Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right)\middle| J=\hat{J}, \hat{\tau}_j^* = t_j, \hat{\nu}_j^{(1)} = v_j \text{ for all } j\right] \\
= \quad &\mathbb{P}\left[\left|\underset{t\in[t_j\pm\hat{d}_j]-\tau_j^{**}}{\arg\min}\ \hat{X}_j(t+\tau_j^{**})\right| \leq Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right)\middle| J=\hat{J}, \hat{\tau}_j^* = t_j, \hat{\nu}_j^{(1)} = v_j \text{ for all } j\right]
\end{aligned}$$

$$\begin{aligned}
&\geq \quad \mathbb{P}\left[\left|\underset{t\in[t_j\pm\hat{d}_j]-\tau_j^{**}}{\arg\min}\ \frac{1}{\sigma}X_j'(t)\right| \leq Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right)\right] - C_1(|\underline{\Delta}|/2)|\hat{D}_j| \\
&\geq \quad \mathbb{P}\left[\left|\underset{t\in[t_j\pm\hat{d}_j]-\tau_j^{**}}{\arg\min}\ \frac{1}{\sigma}X_j'(t)\right| \leq Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right)\right] - C_1(|\underline{\Delta}|/2)\rho_N \qquad \text{(A.105)}
\end{aligned}$$

for all large $N$, where $C_1(|\underline{\Delta}|/2)$ is some expression independent of $N$ and $j$.

Next, use Lemma 11 to obtain

$$\begin{aligned}
&\mathbb{P}\left[\left|\underset{t\in[t_j\pm\hat{d}_j]-\tau_j^{**}}{\arg\min}\ \frac{1}{\sigma}X_j'(t)\right| \leq Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right)\right] \\
\geq \quad &\mathbb{P}\left[\left|\underset{t\in[t_j\pm\hat{d}_j]-\tau_j^{**}}{\arg\min}\ X_{(|\Delta_j|+2\rho_N)/\sigma}(t)\right| \leq Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right)\right] \qquad \text{(A.106)}
\end{aligned}$$

We could also use Lemma 6 twice to obtain

$$\begin{aligned}
&\mathbb{P}\left[\left|\underset{t\in[t_j\pm\hat{d}_j]-\tau_j^{**}}{\arg\min}\ X_{(|\Delta_j|+2\rho_N)/\sigma}(t)\right| \leq Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right)\right] \\
\geq \quad &P_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right) - C_2(\underline{\Delta})\exp(-C_3(\underline{\Delta})\Delta_{N^*}^*) \qquad \text{(A.107)}
\end{aligned}$$

for some positive expressions $C_2(\underline{\Delta})$ and $C_3(\underline{\Delta})$ independent of $N$ and $j$. By combining these inequalities, we come to the conclusion that

$$\begin{aligned}
&\mathbb{P}\left[|\hat{\tau}_j^{re} - \tau_j^{**}| \leq Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right)\middle| J=\hat{J}, \hat{\tau}_j^* = t_j, \hat{\nu}_j^{(1)} = v_j \text{ for all } j\right] \\
\geq \quad &P_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right) - C_1(|\underline{\Delta}|/2)\rho_N - C_2(\underline{\Delta})\exp(-C_3(\underline{\Delta})\delta_{N^*}^*) \qquad \text{(A.108)}
\end{aligned}$$

Therefore

$$\begin{aligned}
& \mathbb{P}\left[\max_{j=1,\ldots,J}|\hat{\tau}_j^{re}-\tau_j^{**}| \le Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right)\ \forall j \Big| J=\hat{J},\hat{\tau}_j^*=t_j,\hat{\nu}_j^{(1)}=v_j\ \forall j\right] \\
= \quad & \mathbb{P}\left[\bigcap_{j=1,\ldots,J}|\hat{\tau}_j^{re}-\tau_j^{**}| \le Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right)\ \forall j \Big| J=\hat{J},\hat{\tau}_j^*=t_j,\hat{\nu}_j^{(1)}=v_j\ \forall j\right] \\
\ge \quad & 1-\sum_{j=1}^{J}\mathbb{P}\left[|\hat{\tau}_j^{re}-\tau_j^{**}| > Q_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right)\Big| J=\hat{J},\hat{\tau}_j^*=t_j,\hat{\nu}_j^{(1)}=v_j \text{ for all } j\right] \\
\ge \quad & 1-J\left[1-P_{(|\Delta_j|+2\rho_N)/\sigma}\left(1-\frac{\alpha_N}{J}\right)+(C_1(|\underline{\Delta}|/2)\rho_N+C_2(\underline{\Delta})\exp(-C_3(\underline{\Delta})\delta_{N^*}^*))\right] \\
& \ge 1-\alpha_N-J\left(C_1(|\underline{\Delta}|/2)\rho_N+C_2(\underline{\Delta})\exp(-C_3(\underline{\Delta})\delta_{N^*}^*)\right) \qquad \text{(A.109)}
\end{aligned}$$

which converges to 1 since $\alpha_N \to 0$, $J\rho_N \to 0$, and $\delta_{N^*}^* > C(N^*)^\eta$ for some positive constant $C$ and $\eta$. □

### A.2.3 Proof of Theorem II.5

In this section we again utilize the $\pi_2$ function defined in (A.61), which is a bijection from the set $\{1,\ldots,N\}-S^{(1)}$ to the set $\{1,\ldots,N-|S^{(1)}|\}$.

*Proof.* Let $S^{(2)}\left(\hat{\tau}_k^{(1)}\right)$ for $k=1,...,\hat{J}$ be the second stage subsamples. As in previous sections, define

$$\mathcal{R}_N := \left\{\hat{J}=J;\ \max_{j=1,\ldots,J}\left|\hat{\tau}_j^{(1)}-\tau_j\right| \le w(N);\ \max_{j=0,\ldots,J}\left|\hat{\nu}_j^{(1)}-\nu_j\right| \le \rho_N\right\} \qquad \text{(A.110)}$$

We also define the following random functions on integers: for $k=1,...,J$, on the event $\mathcal{R}_N$ let

$$\begin{aligned}
\hat{X}_k^{(2)}(d) := \operatorname{sgn}\left(\hat{\nu}_k^{(1)}-\hat{\nu}_{k-1}^{(1)}\right) & \sum_{j\in S^{(2)}(\hat{\tau}_k^{(1)})}\left(Y_j-\frac{\hat{\nu}_k^{(1)}+\hat{\nu}_{k-1}^{(1)}}{2}\right)\cdot \\
& \left(1\left(j\le\pi_2^{-1}(\pi_2(\tau_k^{(2)})+d)\right)-1\left(j\le\tau_k^{(2)}\right)\right) \\
& \text{for } \pi_2^{-1}(\pi_2(\tau_k^{(2)})+d)\in S_k^{(2)} \\
\hat{X}_k^{(2)}(d) := \infty \qquad & \text{otherwise} \qquad \text{(A.111)}
\end{aligned}$$

and on the event $\mathcal{R}_N^C$ let

$$\hat{X}_k^{(2)}(d) := d. \tag{A.112}$$

so that the arg min of $\hat{X}_k^{(2)}(d)$ is $d = -\infty$ on the event $\mathcal{R}_N^C$. Using this definition, for all sufficiently large $N$, the event $\{\hat{J} = J,\ \lambda_2(\tau_k^{(2)}, \hat{\tau}_k^{(2)}) = j_k \text{ for } k = 1, ..., J\} \cap \mathcal{R}_N$ is equivalent to the event

$$\left\{ \arg\min_{d \in \mathbb{Z}} \hat{X}_k^{(2)}(d) = j_k \text{ for } k = 1, ..., J \right\}.$$

Further, for any integer $M > 0$, we can easily obtain convergence properties by restricting the function $\hat{X}_k^{(2)}$ to the set $\{-M, -(M-1), ..., M\}$. For sufficiently large $N$, when event $\mathcal{R}_N$ occurs, we have for any $d \in \{-M, ..., M\}$, and for all $k = 1, ..., J$,

$$\begin{aligned}
\hat{X}_k^{(2)}(d) &= \operatorname{sgn}(\Delta_k) \sum_{j \in S^{(2)}(\hat{\tau}_k^{(1)})} \left( Y_j - \frac{\hat{\nu}_k^{(1)} + \hat{\nu}_{k-1}^{(1)}}{2} \right) \left( 1\left( j \le \pi_2^{-1}(\pi_2(\tau_k^{(2)}) + d) \right) - 1\left( j \le \tau_k^{(2)} \right) \right) \\
&= \begin{cases}
\operatorname{sgn}(\Delta_k) \sum\limits_{j=\pi_2(\tau_k^{(2)})+1}^{\pi_2(\tau_k^{(2)})+d} \left( Y_{\pi^{-1}(j)} - \nu_k + \frac{\nu_k - \nu_{k-1}}{2} + \frac{1}{2}\left( \nu_k - \hat{\nu}_k^{(1)} + \nu_{k-1} - \hat{\nu}_{k-1}^{(1)} \right) \right) & \text{for } d > 0 \\
0 & \text{for } d = 0 \\
-\operatorname{sgn}(\Delta_k) \sum\limits_{j=\pi_2(\tau_k^{(2)})+d+1}^{\pi_2(\tau_k^{(2)})} \left( Y_{\pi_2^{-1}(j)} - \nu_{k-1} + \frac{\nu_{k-1} - \nu_k}{2} + \frac{1}{2}\left( \nu_k - \hat{\nu}_k^{(1)} + \nu_{k-1} - \hat{\nu}_{k-1}^{(1)} \right) \right) & \text{for } d < 0
\end{cases} \\
&= \begin{cases}
\frac{d|\Delta_k|}{2} + \operatorname{sgn}(\Delta_k) \left( \sum\limits_{j=\pi_2(\tau_k^{(2)})+1}^{\pi_2(\tau_k^{(2)})+d} \varepsilon_{\pi_2^{-1}(j)} + \frac{d}{2}\left( \nu_k - \hat{\nu}_k^{(1)} + \nu_{k-1} - \hat{\nu}_{k-1}^{(1)} \right) \right) & \text{for } d > 0 \\
0 & \text{for } d = 0 \\
-d\frac{|\Delta_k|}{2} - \operatorname{sgn}(\Delta_k) \left( \sum\limits_{j=\pi_2(\tau_k^{(2)})+d+1}^{\pi_2(\tau_k^{(2)})} \varepsilon_{\pi_2^{-1}(j)} + \frac{d}{2}\left( \nu_k - \hat{\nu}_k^{(1)} + \nu_{k-1} - \hat{\nu}_{k-1}^{(1)} \right) \right) & \text{for } d < 0
\end{cases}
\end{aligned} \tag{A.113}$$

Because $\max\limits_{i=1,...,J} |\hat{\nu}_i^{(1)} - \nu_o| \le \rho_N$ under $\mathcal{R}_N$, this gives the uniform bound

$$\left| \frac{d}{2\sigma} \left( \nu_k - \hat{\nu}_k^{(1)} + \nu_{k-1} - \hat{\nu}_{k-1}^{(1)} \right) \right| \le \frac{1}{2} M \rho_N. \tag{A.114}$$

for all $k$ and $d$. The right side of the above inequality converges to 0 since $\rho_N \to 0$, and because all of this occurs with probability $\mathbb{P}[\mathcal{R}_N] \to 1$, this shows that the $\hat{X}_k^{(2)}(d)$'s all jointly converge. Specifically, let $\varepsilon_{i,j}^*$ for $i = 1, .., J$ and $j \in \mathbb{Z}$ be random variables with the same distribution as the error terms of the data sequence, and define the random walks

$$X_{\Delta_k}^*(d) = \begin{cases} d\frac{|\Delta_k|}{2} + \text{sgn}(\Delta_k) \sum_{j=1}^{d} \varepsilon_{k,j}^* & \text{for } d > 0 \\ 0 & \text{for } d = 0 \\ -d\frac{|\Delta_k|}{2} - \text{sgn}(\Delta_k) \sum_{j=d+1}^{0} \varepsilon_{k,j}^* & \text{for } d < 0 \end{cases} \tag{A.115}$$

(note that if the error terms are iid $N(0, \sigma^2)$, the random walk $X_{\Delta_k}^*(d)$ has precisely the same distribution as the random walk $\sigma X_{\Delta_k/\sigma}(d)$). We have the joint weak convergence

$$\begin{pmatrix} \hat{X}_1^{(2)}(-M), & \hat{X}_1^{(2)}(-M+1), & \ldots, & \hat{X}_1^{(2)}(M), \\ \hat{X}_2^{(2)}(-M), & \hat{X}_2^{(2)}(-M+1), & \ldots, & \hat{X}_2^{(2)}(M), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ \hat{X}_J^{(2)}(-M), & \hat{X}_J^{(2)}(-M+1), & \ldots, & \hat{X}_J^{(2)}(M) \end{pmatrix} \quad \Rightarrow$$
$$\begin{pmatrix} X_{\Delta_1}^*(-M), & X_{\Delta_1}^*(-M+1), & \ldots, & X_{\Delta_1}^*(M), \\ X_{\Delta_2}^*(-M), & X_{\Delta_2}^*(-M+1), & \ldots, & X_{\Delta_2}^*(M), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ X_{\Delta_J}^*(-M), & X_{\Delta_J}^*(-M+1), & \ldots, & X_{\Delta_J}^*(M) \end{pmatrix} \tag{A.116}$$

Define $L_{\Delta_k}^* := \arg\min_{j \in \mathbb{Z}} X_{\Delta_k}^*(j)$, and $L_{\Delta_k}^{*(M)} := \arg\min_{|j| \leq M} X_{\Delta_k}(j)$ for $k = 1, \ldots, J$ (note that if the error terms of the data sequence is $N(0, 1)$, $L_{\Delta_k}^*$ has the same distribution

as $L_{\Delta_k/\sigma}$). We have the joint weak convergence

$$\left(\underset{|j|\leq M}{\arg\max}\hat{X}_1^{(2)}(d),\ldots,\underset{|j|\leq M}{\arg\max}\hat{X}_J^{(2)}(d)\right) \quad \Rightarrow \quad \left(L_{\Delta_1}^{*(M)},\ldots,L_{\Delta_J}^{*(M)}\right) \qquad \text{(A.117)}$$

by the continuous mapping theorem, because arg max is a continuous function on $\mathbb{R}^{2M+1}$ (except when at least two of the coordinates are equal, which has probability 0 if the error terms have densities).

Next, we establish that $\mathbb{P}\left[\hat{J}=J,\ \lambda_2(\tau_k^{(2)},\hat{\tau}_k^{(2)})=j_k \text{ for } k=1,...,J\right]$ converges to the product of $\mathbb{P}[L_{\Delta_k}^*=j_k]$ (which equal $\mathbb{P}[L_{\Delta_k/\sigma}=j_k]$ for $N(0,\sigma^2)$ error terms) for $k=1,...,J$. We will do this by showing for any fixed $\epsilon>0$, the absolute difference between the two is smaller than $\epsilon$ for all large $N$. As in the proof of the single change point problem, this is accomplished through three main inequalities.

**First Inequality**: From the result of Theorem II.4, and by the fact that $\mathbb{P}[\mathcal{R}_N]\to 1$, we can find an integer $K_0$ greater than $\max_k|j_k|$, such that for any $K_1\geq K_0$, we have for sufficiently large $N$

$$\mathbb{P}\left[\hat{J}=J,\ \max_{k=1,...,J}\left|\hat{\tau}_k^{(2)}-\tau_k\right|\leq(2K_1+2),\ \mathcal{R}_N\right]\geq 1-\frac{\epsilon}{4} \qquad \text{(A.118)}$$

For all sufficiently large $N$, $\left\{\hat{J}=J,\ \max_{k=1,...,J}\left|\hat{\tau}_k^{(2)}-\tau_k\right|\leq(2K_1+2)\right\}\cap\mathcal{R}_N$ would mean

$$\left\{\hat{J}=J,\ \max_{k=1,..,J}\left|\lambda_2\left(\tau_k^{(2)},\hat{\tau}_k^{(2)}\right)\right|\leq K_1\right\}\cap\mathcal{R}_N$$

and hence

$$\begin{aligned} 1-\frac{\epsilon}{3} &\leq \mathbb{P}\left[\hat{J}=J,\ \min_{k=1,..,J}\left|\lambda_2\left(\tau_k^{(2)},\hat{\tau}_k^{(2)}\right)\right|\leq K_1,\ \mathcal{R}_N\right] \\ &= \mathbb{P}\left[\hat{J}=J;\ \max_{k=1,...,J}\left|\arg\min_{d\in\mathbb{Z}}\hat{X}_k^{(2)}(d)\right|\leq K_1\right] \end{aligned} \tag{A.119}$$

Now

$$\begin{aligned} & \arg\min_{d\in\mathbb{Z}}\hat{X}_k^{(2)}(d)=j_k \text{ for } k=1,...,J \\ \longleftrightarrow\ & \arg\min_{|d|\leq K_1}\hat{X}_k^{(2)}(d)=j_k \text{ for } k=1,...,J, \text{ and } \max_{k=1,...,J}\left|\arg\min_{d\in\mathbb{Z}}\hat{X}_k^{(2)}(d)\right|\leq K_1 \end{aligned} \tag{A.120}$$

With steps very similar to those used in (A.45), it can be shown that

$$\begin{aligned} & \left|\mathbb{P}\left[\arg\min_{d\in\mathbb{Z}}\hat{X}_k^{(2)}(d)=j_k \text{ for } k=1,...,J\right]\right. \\ & \left.-\mathbb{P}\left[\arg\min_{|d|\leq K_1}\hat{X}_k^{(2)}(d)=j_k \text{ for } k=1,...,J\right]\right| \\ \leq\ & \mathbb{P}\left[\max_{k=1,...,J}\left|\arg\min_{d\in\mathbb{Z}}\hat{X}_k^{(2)}(d)\right|>K_1\right] \\ & \leq \epsilon/4 \end{aligned} \tag{A.121}$$

**Second Inequality** We can find some integer $K_2 > K_0$ such that

$$\mathbb{P}\left[\max_{k=1,..,J}|L^*_{\Delta_k}|\leq K_2\right]\geq 1-\frac{\epsilon}{4} \tag{A.122}$$

Now $L^*_{\Delta_k}=j_k$ for $k=1,...,J$ if and only if both $L^{*(K_2)}_{\Delta_k}=j_k$ for $k=1,...,J$ and

$\max_k |L_{\Delta_k/\sigma}| \leq K_2$. With steps very similar to those in (A.41), we have

$$\begin{aligned} & \left| \mathbb{P}\left[L^*_{\Delta_k} = j_k \text{ for } k = 1, ..., J\right] - \mathbb{P}\left[L^{*(K_2)}_{\Delta_k} = j_k \text{ for } k = 1, ..., J\right] \right| && \text{(A.123)} \\ \leq \; & \mathbb{P}\left[\max_{k=1,..,J} |L^*_{\Delta_k}| > K_2\right] \\ \leq \; & \epsilon/4 && \text{(A.124)} \end{aligned}$$

**Third Inequality** By weak convergence, we have

$$\left| \mathbb{P}\left[L^{*(K_2)}_{\Delta_k} = j_k \text{ for } k = 1, ..., J\right] - \mathbb{P}\left[\arg\min_{|d| \leq K_2} \hat{X}^{(2)}_k(d) = j_k \text{ for } k = 1, ..., J\right] \right| \leq \frac{\epsilon}{4} \tag{A.125}$$

for all sufficiently large $N$.

Combining the inequalities in (A.121), (A.123), and (A.125) will give

$$\begin{aligned} & \left| \mathbb{P}\left[L^*_{\Delta_k} = j_k \text{ for } k = 1, ..., J\right] - \mathbb{P}\left[\hat{J} = J,\, \lambda_2\left(\tau^{(2)}_k, \hat{\tau}^{(2)}_k\right) = j_k \text{ for } k = 1, .., J,\, \mathcal{R}_N\right] \right| \\ = \; & \left| \mathbb{P}\left[L^*_{\Delta_k} = j_k \text{ for } k = 1, ..., J\right] - \mathbb{P}\left[\arg\min_{d \in \mathbb{Z}} \hat{X}^{(2)}_k(d) = j_k \text{ for } k = 1, ..., J\right] \right| \\ \leq \; & \left| \mathbb{P}\left[\arg\min_{d \in \mathbb{Z}} \hat{X}^{(2)}_k(d) = j_k \text{ for } k = 1, ..., J\right] - \mathbb{P}\left[\arg\min_{|d| \leq K_2} \hat{X}^{(2)}_k(d) = j_k \text{ for } k = 1, ..., J\right] \right| \\ + \; & \left| \mathbb{P}\left[L^{*(K_2)}_{\Delta_k} = j_k \text{ for } k = 1, ..., J\right] - \mathbb{P}\left[\arg\min_{|d| \leq K_2} \hat{X}^{(2)}_k(d) = j_k \text{ for } k = 1, ..., J\right] \right| \\ + \; & \left| \mathbb{P}\left[L^*_{\Delta_k} = j_k \text{ for } k = 1, ..., J\right] - \mathbb{P}\left[L^{*(K_2)}_{\Delta_k} = j_k \text{ for } k = 1, ..., J\right] \right| \\ & \leq 3\epsilon/4 && \text{(A.126)} \end{aligned}$$

Additionally, for sufficiently large $N$ we have $\mathbb{P}[\text{not } \mathcal{R}_N] < \epsilon/4$ and hence

$$\begin{aligned} & \Bigg| \mathbb{P}\left[\hat{J} = J,\, \lambda_2\left(\tau^{(2)}_k, \hat{\tau}^{(2)}_k\right) = j_k \text{ for } k = 1, .., J,\, \mathcal{R}_N\right] \\ - \; & \mathbb{P}\left[\hat{J} = J,\, \lambda_2\left(\tau^{(2)}_k, \hat{\tau}^{(2)}_k\right) = j_k \text{ for } k = 1, .., J\right] \Bigg| \\ < \; & \epsilon/4 && \text{(A.127)} \end{aligned}$$

and hence

$$\left| \mathbb{P}\left[L^*_{\Delta_k} = j_k \text{ for } k = 1, ..., J\right] \right.$$
$$\left. - \quad \mathbb{P}\left[\hat{J} = J,\, \lambda_2\left(\tau_k^{(2)}, \hat{\tau}_k^{(2)}\right) = j_k \text{ for } k = 1, .., J\right] \right| < \epsilon \tag{A.128}$$

for all sufficiently large $N$. □

### A.2.4 Proof of Lemma 1

In order to prove Lemma 1, we will rely on the following result:

**Lemma 4.** *(i) Suppose that we have an estimation scheme which when applied onto $Z_1, \ldots, Z_{N^*}$ gives estimates $\hat{J}$ for $J$ and $\left(\hat{\tau}_1^*, \cdots, \hat{\tau}_{\hat{J}}^*\right)$ for $(\tau_1^*, \cdots, \tau_J^*)$ such that*

$$\mathbb{P}\left[\hat{J} = J; \quad \max_{i=1,\cdots,J} |\hat{\tau}_j^* - \tau_j^*| \leq w^*(N^*)\right] \geq 1 - B^*_{N^*} \tag{A.129}$$

*for some sequences $w^*(N^*)$ and $B_{N^*}$, with $w^*(N^*) = o(\delta^*_{N^*})$ and $B_{N^*} \to 0$. Then, for any positive sequence $\{\rho^*_{N^*}\}$ such that $\frac{w^*(N^*)}{\delta^*_{N^*}} = o(\rho^*_{N^*})$, there exist constants $C_1$ and $C_2$, where*

$$\mathbb{P}\left[\hat{J} = J; \quad |\hat{\nu}_i^* - \nu_i^*| \geq \rho^*_{N^*}\right] \leq B_{N^*} + C_1 w^*(N^*) \frac{\exp\left[-C_2 \delta^*_{N^*} \rho^{*2}_{N^*}\right]}{\sqrt{\delta^*_{N^*}} \rho^*_{N^*}} \tag{A.130}$$

*for all $i = 1, \cdots, J$, when $N^*$ is sufficiently large.*

*(ii) Moreover, as a consequence of part (i),*

$$\mathbb{P}\left[\hat{J} = J; \quad \max_{i=0,\cdots,J} |\hat{\nu}_i^* - \nu_i^*| < \rho^*_{N^*}\right] \geq 1 - \left(\frac{N^*}{\delta^*_{N^*}} + 2\right) B_{N^*}$$
$$- C_1 \left(\frac{N^*}{\delta^*_{N^*}} + 1\right) w^*(N^*) \frac{\exp[-C_2 \delta^*_{N^*} \rho^{*2}_{N^*}]}{\sqrt{\delta^*_{N^*}} \rho^*_{N^*}}\,. \tag{A.131}$$

*It follows that in addition to the conditions in (i), if, furthermore, $N^* B_{N^*} / \delta^*_{N^*} \to 0$*

*and* $(N^* w^*(N^*)/\delta_{N^*}^{*3/2}\rho_{N^*}^*) = o(\exp[C_2\delta_{N^*}^*\rho_{N^*}^{*2}])$*, then the probability in* (A.*131) goes to 1. The* $\hat{\nu}_i^*$*'s are simultaneously consistent if* $\rho_{N^*}^*$ *also converges to 0.*

*Proof.* See Section A.2.5 □

In order to prove Lemma 1, it is sufficient to find a sequence $\rho_{N^*}^* \to 0$ such that $\rho_{N^*}^*$ satisfies all the conditions of Lemma 4 and $J\rho_{N^*}^* \to 0$ . The proof will proceed in such a fashion.

*Proof.* We start off by defining some notations. Since $\delta_N \geq CN^{1-\Xi}$ by (M3), we must have $J \leq C'N^{\Lambda}$ for some $C' > 0$ and $\Lambda \in [0, \Xi]$. Using this notation, we will show that by setting $\rho_{N^*}^* = (N^*)^{\theta}$ where $\theta$ is chosen to be any value in $\Big((3\Xi/\gamma - 1) \vee (-3/8), -\frac{\Lambda}{\gamma}\Big)$, we will have a $\rho_{N^*}^* \to 0$ which satisfies the conditions of Lemma 4 and $J\rho_{N^*}^* \to 0$.

We must verify that $\Big((3\Xi/\gamma - 1) \vee (-3/8), -\frac{\Lambda}{\gamma}\Big)$ is a nonempty set by showing that both $(3\Xi/\gamma - 1)$ and $-3/8$ are strictly smaller than $-\frac{\Lambda}{\gamma}$. First, due to condition (M7 (BinSeg)) we know that $\Xi/\gamma < 1/4$, and therefore

$$\begin{aligned} & \frac{3\Xi}{\gamma} + \frac{\Lambda}{\gamma} \leq \frac{4\Xi}{\gamma} < 1 \\ \to \quad & \frac{3\Xi}{\gamma} - 1 < -\frac{\Lambda}{\gamma}, \end{aligned} \tag{A.132}$$

and additionally,

$$-\frac{3}{8} < -\frac{1}{7} \leq -\frac{\Xi}{\gamma} \leq -\frac{\Lambda}{\gamma}. \tag{A.133}$$

Therefore, it is possible to choose some value of $\theta$ within the set $\Big((3\Xi/\gamma - 1) \vee (-3/8), -\frac{\Lambda}{\gamma}\Big)$.

To verify that $J\rho^*_{N^*} \to 0$, we first note

$$J\rho^*_{N^*} \lesssim N^{\Lambda}(N^*)^{\theta} \lesssim N^{\Lambda+\gamma\theta}.$$

The rightmost term goes to 0 because $\theta < \Lambda/\gamma$.

To show that the BinSeg estimators satisfy the conditions of Lemma 4, we proceed as follows. First note that $\delta^*_{N^*} \geq C_1(N^*)^{1-\Xi/\gamma}$ for some $C_1 > 0$, and where $\Xi/\gamma < 1/4$. For some positive constant $C_2$, set $w^*(N^*) = C_2 E_{N^*} = C_2 \left(\frac{N^*}{\delta^*_{N^*}}\right)^2 \log(N^*)$. Then, there is a positive constant $C_4$ such that $w^*(N^*) = (C_4 + o(1))(N^*)^{2\Xi/\gamma} \log(N^*)$. Set $B_{N^*} = C_5/N^*$; this would mean

- since $(N^*)^{2\Xi/\gamma} \log(N^*) = o((N^*)^{1-\Xi/\gamma})$ this does allow $w^*(N^*) = o(\delta^*_{N^*})$ to be satisfied;
- $\frac{N^*}{\delta^*_{N^*}} B_{N^*} = \frac{C_5}{\delta^*_{N^*}} \to 0$;
- since $3\Xi/\gamma - 1 < 0$, and because $\rho^*_{N^*} = (N^*)^{\theta}$ for some $\theta$ satisfying $(3\Xi/\gamma - 1) \vee (-3/8) < \theta < 0$, this $\rho^*_{N^*} \to 0$ and satisfies:
  - $\frac{w^*(N^*)}{\delta^*_{N^*}} \leq C_6(N^*)^{3\Xi/\gamma - 1} \log(N^*)$ for some $C_6 > 0$; latter expression is $o(\rho^*_{N^*})$;
  - $N^* w^*(N^*)/({\delta^*_{N^*}}^{3/2} \rho^*_{N^*}) \leq C_7(N^*)^{((7\Xi/\gamma - 1)/2) - \theta}$ and $\exp[C_2 \delta^*_{N^*}(\rho^*_{N^*})^2] \geq \exp[C_8(N^*)^{1-\Xi/\gamma+2\theta}]$, for some positive $C_7$, $C_8$; as $1 - \Xi/\gamma + 2\theta > 3/4 + 2\theta > 0$ it follows that $N^* w^*(N^*)/{\delta^*_{N^*}}^{3/2} \rho^*_{N^*} = o(\exp[C_2 \delta^*_{N^*}(\rho^*_{N^*})^2])$.

Therefore, all conditions of Lemma 4 for a sequence $\rho^*_{N^*}$ tending to 0 are satisfied. Next, combining the results of Theorem II.6 and Lemma 4, we establish the simultaneous consistency of $\hat{J}$, the $\hat{\tau}_i$'s, and the $\hat{\nu}_i$'s. Specifically, under conditions (M1) to

(M6 (BinSeg)), we could combine the two limit results

$$\begin{aligned}
&\mathbb{P}\left[\hat{J} = J; \quad \max_{i=1,\ldots,J} |\hat{\tau}_i^* - \tau_i^*| \leq CE_{N^*}\right] \to 1 \\
&\mathbb{P}\left[\hat{J} = J; \quad \max_{i=0,\ldots,J} |\hat{\nu}_i^* - \nu_i^*| \leq \rho_{N^*}^*\right] \to 1 \\
&\text{for any } \rho_{N^*}^* = (N^*)^\theta, \text{ where } \theta \in \left((3\Xi/\gamma - 1) \vee \left(-\frac{3}{8}\right)\right)
\end{aligned} \tag{A.134}$$

to get the following through the Bonferroni inequality:

$$\mathbb{P}\left[\hat{J} = J; \quad \max_{i=1,\ldots,J} |\hat{\tau}_i^* - \tau_i^*| \leq CE_{N^*}; \quad \max_{i=0,\ldots,J} |\hat{\nu}_i^* - \nu_i^*| \leq \rho_{N^*}^*\right] \to 1 \tag{A.135}$$

as $N^* \to \infty$. □

### A.2.5 Proof For Lemma 4

*Proof.* First we focus on part (i). Consider the separate cases of $\tau_i$ for $i = 0$ or $i = J$ (case 1), and $1 \leq i < J$ (case 2). We know

$$\mathbb{P}\Big[\hat{J} = J; \quad \max_{j=1,\ldots,J} |\hat{\tau}_j^* - \tau_j^*| \leq w^*(N^*)\Big] \geq 1 - B_{N^*} \tag{A.136}$$

for some sequences $w^*(N^*)$ and $B_{N^*}$ which are $o(\delta_{N^*}^*)$, and $o(1)$, respectively. Also as in the statement of the theorem, assume there is some sequence $\rho_{N^*}^*$ such that $w^*(N^*)/\delta_{N^*}^* = o(\rho_{N^*})$. For the rest of the proof assume $N^*$ is large enough so that

- $1 < w^*(N^*) < \frac{\delta_{N^*}^*}{6}$
- $6\frac{w^*(N^*)}{\delta_{N^*}^*}\bar{\theta} < \frac{\rho_{N^*}^*}{2}$
- $\delta_{N^*}^* > 3$

**Case 1**: For $i = 0$, $\hat{\nu}_0^*$ is the average of all $Z_t$'s where $t$ lies between 1 and $\hat{\tau}_1^*$, inclusive. We have

$$\begin{aligned}
& \mathbb{P}\left[\hat{J} = J; \quad |\hat{\nu}_0^* - \nu_0^*| \geq \rho_{N^*}^*\right] \\
\leq \quad & \mathbb{P}\left[\hat{J} = J; \quad |\hat{\tau}_1^* - \tau_1^*| > w^*(N^*)\right] \\
& + \mathbb{P}\left[\hat{J} = J; \quad |\hat{\nu}_0^* - \nu_0^*| \geq \rho_{N^*}^*; \quad |\hat{\tau}_1^* - \tau_1^*| \leq w^*(N^*)\right] \\
\leq \quad & B_{N^*} + \sum_{\tau:\, |\tau - \tau_1^*| \leq w^*(N^*)} \mathbb{P}\left[\hat{J} = J; \quad |\hat{\nu}_0^* - \nu_0^*| \geq \rho_{N^*}^*; \quad \hat{\tau}_1^* = \tau\right] \\
\leq \quad & B_{N^*} + \sum_{\tau:\, |\tau - \tau_1^*| \leq w^*(N^*)} \mathbb{P}\left[\hat{J} = J; \quad \hat{\tau}_1^* = \tau; \quad \left|\frac{1}{\tau}\sum_{j=1}^{\tau} Z_j - \nu_0^*\right| \geq \rho_{N^*}^*\right] \\
& \leq B_{N^*} + \sum_{\tau:\, |\tau - \tau_1^*| \leq w^*(N^*)} \mathbb{P}\left[\left|\frac{1}{\tau}\sum_{j=1}^{\tau} (Z_j - \nu_0^*)\right| \geq \rho_{N^*}^*\right] \qquad \text{(A.137)}
\end{aligned}$$

For all $\tau_1 - w^*(N^*) \leq \tau \leq \tau_1^*$, we have $\frac{1}{\tau}\sum_{j=1}^{\tau}(Z_j - \nu_0^*) \sim N(0, \sigma^2/\tau)$, and hence

$$\begin{aligned}
\mathbb{P}\left[\left|\frac{1}{\tau}\sum_{j=1}^{\tau}(Z_j - \tau_0)\right| \geq \rho_{N^*}^*\right] &= 2\left(1 - \Phi(\sqrt{\tau}\rho_{N^*}^*)\right) \\
&\leq 2\frac{\phi(\sqrt{\tau}\rho_{N^*}^*/\sigma)}{\sqrt{\tau}\rho_{N^*}^*/\sigma} \\
&\leq \sqrt{\frac{2}{\pi}} \cdot \frac{\exp\left(-(\tau_1 - w^*(N^*))(\rho_{N^*}^*)^2/(2\sigma^2)\right)}{\sqrt{\tau_1 - w^*(N^*)}\rho_{N^*}^*/\sigma} \\
&\leq \frac{2\sigma}{\sqrt{\pi}} \cdot \frac{\exp\left(-\delta_{N^*}^*(\rho_{N^*}^*)^2/(4\sigma^2)\right)}{\sqrt{\delta_{N^*}^*}\rho_{N^*}^*} \\
&\quad \text{(by } \tau_1 w^*(N^*) > \delta_{N^*}^*/2) \qquad \text{(A.138)}
\end{aligned}$$

For all $\tau_1^* < \tau \leq \tau_1^* + w^*(N^*)$ we have $\frac{1}{\tau}\sum_{j=1}^{\tau}(Z_j - \nu_0^*) \sim N\left(\frac{\tau - \tau_1^*}{\tau}(\nu_1^* - \nu_0^*), \frac{\sigma^2}{\tau}\right)$. Because

$$\left|\frac{\tau - \tau_1^*}{\tau}(\nu_1^* - \nu_0^*)\right| \leq \frac{w^*(N^*)}{\delta_{N^*}^*}(2\bar{\theta}) \leq \frac{\rho_{N^*}^*}{2}, \qquad \text{(A.139)}$$

the magnitude of the z-score of both $\pm\rho^*_{N^*}$ for the $N\left(\frac{\tau-\tau_1^*}{\tau}(\nu_1^*-\nu_0^*), \frac{\sigma^2}{\tau}\right)$ distribution is at least $\frac{\rho^*_{N^*}\sqrt{\tau}}{2\sigma}$, and hence

$$\begin{aligned}
\mathbb{P}\left[\left|\frac{1}{\tau}\sum_{j=1}^{\tau}(Z_j-\tau_0^*)\right| \geq \rho^*_{N^*}\right] &\leq 2\left(1-\Phi\left(\frac{\rho^*_{N^*}\sqrt{\tau}}{2}\right)\right) \\
&\leq 2\frac{\phi\left(\frac{\rho^*_{N^*}\sqrt{\tau}}{2\sigma}\right)}{\frac{\rho^*_{N^*}}{2\sigma}\sqrt{\tau}/\sigma} \\
&\leq \frac{2\sigma\sqrt{2}}{\sqrt{\pi}}\cdot\frac{\exp\left(-\delta^*_{N^*}(\rho^*_{N^*})^2/(8\sigma^2)\right)}{\sqrt{\delta^*_{N^*}}\rho^*_{N^*}} \qquad \text{(A.140)}
\end{aligned}$$

Therefore, the expression (A.138) can be bounded from above by

$$\begin{aligned}
&B_{N^*}+(w^*(N^*)+1)\frac{2\sigma}{\sqrt{\pi}}\cdot\frac{\exp\Big(-\delta^*_{N^*}(\rho^*_{N^*})^2/(4\sigma^2)\Big)}{\sqrt{\delta^*_{N^*}}\gamma_{N^*}}+ \\
&w^*(N^*)\frac{2\sigma\sqrt{2}}{\sqrt{\pi}}\cdot\frac{\exp\left(-\delta^*_{N^*}(\rho^*_{N^*})^2/(8\sigma^2)\right)}{\sqrt{\delta^*_{N^*}}\rho^*_{N^*}} \\
&\qquad\leq B_{N^*}+\frac{6\sigma\sqrt{2}}{\sqrt{\pi}}\cdot\frac{w^*(N^*)\exp\left(-\delta^*_{N^*}(\rho^*_{N^*})^2/(8\sigma^2)\right)}{\sqrt{\delta^*_{N^*}}\rho^*_{N^*}}.
\end{aligned} \qquad \text{(A.141)}$$

For $i=J$, a very similar argument will bound $\mathbb{P}\left[\hat{J}=J; \quad |\hat{\nu}_J^*-\nu_J^*|\geq\rho^*_{N^*}\right]$ by the same expression in (A.141).

**Case 2**: The procedure for this case will be similar to the steps for Case 1, but there are a few modifications. For $0<i<J$, $\hat{\nu}_i^*$ is the average of all $Z_t$'s for $\hat{\tau}_i^*<t\leq\hat{\tau}_{i+1}^*$. For the following part we re-write this average by considering the midpoint $\tau_i^{*(m)}:=\frac{\lceil\tau_i^*+\tau_{i+1}^*\rceil}{2}$ where $1<i<J$.

In the case where $\hat{\tau}_i^*$ and $\hat{\tau}_{i+1}^*$ are within $\delta^*_{N^*}/3$ (which is less than $|\tau_{i+1}^*-\tau_i^*|/3$) of $\tau_i^*$ and $\tau_{i+1}^*$ respectively, we have $\hat{\tau}_i^*<\tau_i^{*(m)}<\hat{\tau}_{i+1}^*$, and hence we can bound $|\hat{\nu}_i^*-\nu_i^*|$

by

$$\begin{aligned}
&\left|\frac{1}{\hat{\tau}_{i+1}^{*}-\hat{\tau}_{i}^{*}}\sum_{j=\hat{\tau}_{i}^{*}+1}^{\hat{\tau}_{i+1}^{*}}(Z_j-\nu_i^{*})\right| \\
&=\left|\frac{\tau_i^{*(m)}-\hat{\tau}_i^{*}}{\hat{\tau}_{i+1}^{*}-\hat{\tau}_{i}^{*}}\left(\frac{1}{\tau_i^{*(m)}-\hat{\tau}_i^{*}}\sum_{j=\hat{\tau}_{i}^{*}+1}^{\tau_i^{*(m)}}(Z_j-\nu_i^{*})\right)\right. \\
&\left.+\frac{\hat{\tau}_{i+1}^{*}-\tau_i^{*(m)}}{\hat{\tau}_{i+1}^{*}-\hat{\tau}_{i}^{*}}\left(\frac{1}{\hat{\tau}_{i+1}^{*}-\tau_i^{*(m)}}\sum_{j=\tau_i^{*(m)}+1}^{\hat{\tau}_{i+1}^{*}}(Z_j-\nu_i^{*})\right)\right| \\
&\leq\frac{\tau_i^{*(m)}-\hat{\tau}_i^{*}}{\hat{\tau}_{i+1}^{*}-\hat{\tau}_{i}^{*}}\left|\frac{1}{\tau_i^{*(m)}-\hat{\tau}_i^{*}}\sum_{j=\hat{\tau}_{i}^{*}+1}^{\tau_i^{*(m)}}(Z_j-\nu_i^{*})\right|+\frac{\hat{\tau}_{i+1}^{*}-\tau_i^{*(m)}}{\hat{\tau}_{i+1}^{*}-\hat{\tau}_{i}^{*}}\left|\frac{1}{\hat{\tau}_{i+1}^{*}-\tau_i^{(m)}}\sum_{j=\tau_i^{(m)}+1}^{\hat{\tau}_{i+1}^{*}}(Z_j-\nu_i^{*})\right|
\end{aligned} \tag{A.142}$$

In order for $|\hat{\nu}_i^{*}-\nu_i^{*}|$ to exceed $\rho_{N^*}^{*}$, at least one of $\left|\frac{1}{\tau_i^{*(m)}-\hat{\tau}_i^{*}}\sum_{j=\hat{\tau}_i^{*}+1}^{\tau_i^{*(m)}}(Z_j-\nu_i^{*})\right|$ or $\left|\frac{1}{\hat{\tau}_{i+1}^{*}-\tau_i^{*(m)}}\sum_{j=\tau_i^{*(m)}+1}^{\hat{\tau}_{i+1}^{*}}(Z_j-\nu_i^{*})\right|$ must exceed $\rho_{N^*}^{*}$, or in other words,

$$\begin{aligned}
&\mathbb{P}\left[\hat{J}=J;\quad|\hat{\nu}_i^{*}-\nu_i^{*}|\geq\rho_{N^*}^{*}\right] \\
&\leq\mathbb{P}\left[\hat{J}=J;\quad|\hat{\tau}_i^{*}-\tau_i^{*}|>w^{*}(N^{*})\quad\text{or}\quad|\hat{\tau}_{i+1}^{*}-\tau_{i+1}^{*}|>w(N)\right]+ \\
&\mathbb{P}\left[\hat{J}=J;\quad|\hat{\tau}_i^{*}-\tau_i^{*}|\leq w^{*}(N^{*});\quad|\hat{\tau}_{i+1}^{*}-\tau_{i+1}^{*}|\leq w^{*}(N^{*});\quad|\hat{\nu}_i^{*}-\nu_i^{*}|\geq\rho_{N^*}^{*}\right] \\
&\leq B_{N^*}+\mathbb{P}\left[\hat{J}=J;\quad|\hat{\tau}_i^{*}-\tau_i^{*}|\leq w^{*}(N^{*});\quad|\hat{\tau}_{i+1}^{*}-\tau_{i+1}^{*}|\leq w^{*}(N^{*});\right. \\
&\left.\left|\frac{1}{\tau_i^{*(m)}-\hat{\tau}_i^{*}}\sum_{j=\hat{\tau}_i^{*}+1}^{\tau_i^{*(m)}}(Z_j-\nu_i^{*})\right|\geq\rho_{N^*}^{*}\right] \\
&+\mathbb{P}\left[\hat{J}=J;\quad|\hat{\tau}_i^{*}-\tau_i^{*}|\leq w^{*}(N^{*});\quad|\hat{\tau}_{i+1}^{*}-\tau_{i+1}^{*}|\leq w^{*}(N^{*});\right. \\
&\left.\left|\frac{1}{\hat{\tau}_{i+1}^{*}-\tau_i^{*(m)}}\sum_{j=\tau_i^{*(m)}+1}^{\hat{\tau}_{i+1}^{*}}(Z_j-\nu_i^{*})\right|\geq\rho_{N^*}^{*}\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq B_{N^*} + \mathbb{P}\left[\hat{J} = J; \quad |\hat{\tau}_i^* - \tau_i^*| \leq w^*(N^*); \quad \left|\frac{1}{\tau_i^{*(m)} - \hat{\tau}_i^*} \sum_{j=\hat{\tau}_i^*+1}^{\tau_i^{*(m)}} (Z_j - \nu_i^*)\right| \geq \rho_{N^*}^*\right] \\
&+ \mathbb{P}\left[\hat{J} = J; \quad |\hat{\tau}_{i+1}^* - \tau_{i+1}^*| \leq w^*(N^*); \quad \left|\frac{1}{\hat{\tau}_{i+1}^* - \tau_i^{*(m)}} \sum_{j=\tau_i^{*(m)}+1}^{\hat{\tau}_{i+1}^*} (Z_j - \nu_i^*)\right| \geq \rho_{N^*}^*\right] \\
&\leq B_{N^*} + \sum_{\tau:\, |\tau - \tau_i^*| \leq w^*(N^*)} \mathbb{P}\left[\left|\frac{1}{\tau_i^{*(m)} - \tau} \sum_{j=\tau+1}^{\tau_i^{*(m)}} (Z_j - \nu_i^*)\right| \geq \rho_{N^*}^*\right] \\
&+ \sum_{\tau:\, |\tau - \tau_{i+1}^*| \leq w^*(N^*)} \mathbb{P}\left[\left|\frac{1}{\tau - \tau_i^{*(m)}} \sum_{j=\tau_i^{*(m)}+1}^{\tau} (Z_j - \nu_i^*)\right| \geq \rho_{N^*}^*\right]
\end{aligned}
\tag{A.143}
$$

Next, we will bound $\mathbb{P}\left[\left|\frac{1}{\tau_i^{*(m)} - \tau} \sum_{j=\tau+1}^{\tau_i^{*(m)}} (Z_j - \nu_i^*)\right| \geq \rho_{N^*}^*\right]$ for each $\tau$ such that $|\tau - \tau_i^*| \leq w^*(N^*)$. For $\tau_i^* \leq \tau \leq \tau_i^* + w^*(N^*)$ we have $\frac{1}{\tau_i^{*(m)} - \tau} \sum_{j=\tau+1}^{\tau_i^{*(m)}} (Z_j - \nu_i^*) \sim N\left(0, \frac{\sigma^2}{\tau_i^{*(m)} - \tau}\right)$, and hence

$$
\begin{aligned}
\mathbb{P}\left[\left|\frac{1}{\tau_i^{*(m)} - \tau} \sum_{j=\tau+1}^{\tau_i^{*(m)}} (Z_j - \nu_i^*)\right| \geq \rho_{N^*}^*\right] &\leq 2\left(1 - \Phi\left(\rho_{N^*}^* \sigma^{-1} \sqrt{\tau_i^{*(m)} - \tau}\right)\right) \\
&\leq \frac{2\sigma\sqrt{3}}{\sqrt{\pi}} \cdot \frac{\exp(-\delta_{N^*}^* (\rho_{N^*}^*)^2/(12\sigma^2))}{\rho_{N^*}^* \sqrt{\delta_{N^*}^*}}
\end{aligned}
\tag{A.144}
$$

where we used the fact that $\tau_i^{*(m)} - \tau > \tau_i^{*(m)} - \tau_i^* - w^*(N^*) > \delta_{N^*}^*/3 - \delta_{N^*}^*/6$.

For $\tau_i^* - w^*(N^*) \leq \tau < \tau_i^*$, we have $\frac{1}{\tau_i^{*(m)} - \tau} \sum_{j=\tau+1}^{\tau_i^{*(m)}} (Z_j - \nu_i^*)$
$\sim N\left(\frac{\tau_i^* - \tau}{\tau_i^{*(m)} - \tau}(\nu_i^* - \nu_{i-1}^*), \frac{\sigma^2}{\tau_i^{*(m)} - \tau}\right)$. The z-scores of $\pm\rho_{N^*}^*$ would have magnitudes

greater than

$$\begin{aligned}
&\sigma^{-1}\sqrt{\tau_i^{*(m)}-\tau}\left(\rho_{N^*}^* - \left|\frac{\tau_i^*-\tau}{\tau_i^{*(m)}-\tau}(\nu_i^*-\nu_{i-1}^*)\right|\right) \\
&\geq \sigma^{-1}\sqrt{\frac{\delta_{N^*}^*}{3}}\left(\rho_{N^*}^* - \frac{w^*(N^*)}{\delta_{N^*}^*/3}(2\bar{\theta})\right) \\
&\geq \sqrt{\frac{\delta_{N^*}^*}{3}}\cdot\frac{\rho_{N^*}^*}{2\sigma} \qquad \text{(A.145)}
\end{aligned}$$

Hence, this gives the probability bound

$$\begin{aligned}
&\mathbb{P}\left[\left|\frac{1}{\tau_i^{*(m)}-\tau}\sum_{j=\tau+1}^{\tau_i^{*(m)}}(Z_j-\nu_i^*)\right| \geq \rho_{N^*}^*\right] \\
\leq\ & 2\left(1-\Phi\left(\sqrt{\frac{\delta_{N^*}^*}{3}}\cdot\frac{\rho_{N^*}^*}{2\sigma}\right)\right) \\
\leq\ & \frac{2\sigma\sqrt{6}}{\sqrt{\pi}}\cdot\frac{\exp\left(-\delta(\gamma_{N^*}^*)^2/(24\sigma^2)\right)}{\rho_{N^*}^*\sqrt{\delta_{N^*}^*}}. \qquad \text{(A.146)}
\end{aligned}$$

Putting together the bounds in (A.144) and (A.146) will give

$$\begin{aligned}
&\sum_{\tau:\,|\tau-\tau_i|\leq w^*(N^*)}\mathbb{P}\left[\left|\frac{1}{\tau_i^{*(m)}-\tau}\sum_{j=\tau+1}^{\tau_i^{*(m)}}(Z_j-\nu_i^*)\right| \geq \rho_{N^*}^*\right] \\
&\leq 3w^*(N^*)\cdot\frac{2\sigma\sqrt{6}}{\sqrt{\pi}}\cdot\frac{\exp\left(-\delta_{N^*}^*(\rho_{N^*}^*)^2/(24\sigma^2)\right)}{\rho_{N^*}^*\sqrt{\delta_{N^*}^*}} \qquad \text{(A.147)}
\end{aligned}$$

In an extremely similar manner, it can be argued that

$$\begin{aligned}
&\sum_{\tau:\,|\tau-\tau_{i+1}^*|\leq w^*(N^*)}\mathbb{P}\left[\left|\frac{1}{\tau-\tau_i^{*(m)}}\sum_{j=\tau_i^{*(m)}+1}^{\tau}(Z_j-\nu_i^*)\right| \geq \rho_{N^*}^*\right] \\
&\leq 3w^*(N^*)\cdot\frac{2\sigma\sqrt{6}}{\sqrt{\pi}}\cdot\frac{\exp\left(-\delta_{N^*}^*(\rho_{N^*}^*)^2/(24\sigma^2)\right)}{\rho_{N^*}^*\sqrt{\delta_{N^*}^*}} \qquad \text{(A.148)}
\end{aligned}$$

Therefore, (A.143) can be bounded by

$$B_{N^*} + \frac{12\sigma\sqrt{6}}{\sqrt{\pi}} \cdot \frac{w^*(N^*)\exp\left(-\delta^*_{N^*}(\rho^*_{N^*})^2/(24\sigma^2)\right)}{\rho^*_{N^*}\sqrt{\delta^*_{N^*}}} \tag{A.149}$$

By taking constants $C_1$ and $C_2$ to be the "worse" of the coefficients in (A.141) and (A.149), which are $\frac{12\sigma\sqrt{6}}{\sqrt{\pi}}$ and $1/(24\sigma^2)$ respectively, we can combine the result of both cases and establish

$$\mathbb{P}\left[\hat{J} = J;\ |\hat{\nu}^*_i - \nu^*_i| \geq \rho^*_{N^*}\right] \leq B_{N^*} + C_1 w^*(N^*)\frac{\exp\left(-C_2\delta^*_{N^*}(\rho^*_{N^*})^2\right)}{\rho^*_{N^*}\sqrt{\delta^*_{N^*}}} \tag{A.150}$$

for all $i = 1, \ldots, J$ □

Using part (i), previously shown, it is straightforward to show part (ii):

*Proof.* The complement of the event $\{\hat{J} = J; \quad \max_{i=0,\ldots,J}|\hat{\nu}_i - \nu_i| < \rho_N\}$ is the event where either $\hat{J} \neq J$ or $\hat{J} = J$ and $|\hat{\nu}_i - \nu_i| \geq \rho_N$ for some $i$. For all sufficiently large $N$ and some positive constants $C_1$ and $C_2$ we have

$$\begin{aligned}
& 1 - \mathbb{P}\left[\hat{J} = J; \quad \max_{i=0,\ldots,J}|\hat{\nu}^*_i - \nu^*_i| < \rho^*_{N^*}\right] \\
\leq\ & \mathbb{P}[\hat{J} \neq J] + \sum_{i=0}^{J}\mathbb{P}\left[\hat{J} = J; \quad |\hat{\nu}^*_i - \nu^*_i| \geq \rho^*_{N^*}\right] \\
\leq\ & B_{N^*} + (J+1)\left(B_{N^*} + C_1 w^*(N^*)\frac{\exp\left[-C_2\delta^*_{N^*}(\rho^*_{N^*})^2\right]}{\sqrt{\delta^*_{N^*}}\rho^*_{N^*}}\right) \\
\leq\ & B_{N^*} + \left(\frac{N^*}{\delta^*_{N^*}} + 1\right)\left(B_{N^*} + C_1 w^*(N^*)\frac{\exp\left[-C_2\delta^*_{N^*}(\rho^*_{N^*})^2\right]}{\sqrt{\delta^*_{N^*}}\rho^*_{N^*}}\right) \\
\to\ & 0
\end{aligned} \tag{A.151}$$

□

### A.2.6 Alternative Proof of Theorem II.4

In this section, we will relaxed one of the conditions used in the previous sections. We will again work with the model presented as (2.11) in the main chapter, assuming conditions (M1), (M2), (M3), and the consistency condition presented in (2.13). Instead of assuming (M4), we instead consider the relaxed condition in which the error terms $\varepsilon_i$'s (for $1 \leq i \leq N$) are iid subGaussian with variance proxy parameter $\sigma^2$ and mean 0, by which we mean they satisfy

$$\mathbb{E}[\exp(s\varepsilon_i)] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right) \tag{A.152}$$

for all $s \in \mathbb{R}$.

In this proof, we again define the parameters

$$\tau_j^{(2)} := \begin{cases} \tau_j - 1 & \text{if } \tau_j \text{ was a first stage subsmaple point} \\ \tau_j & \text{otherwise} \end{cases} \tag{A.153}$$

for $j = 1, \ldots, J$. Also as in previous sections, define

$$S^{(2)}(t) := \{i \in \mathbb{N} : \ |i - t| \leq Kw(N), \quad Y_i \text{ not used in 1st stage subsample}\} \tag{A.154}$$

*Proof.* Define the event

$$\mathcal{R}_N := \left\{ \hat{J} = J; \quad \max_{i=1,\ldots,J} \left| \hat{\tau}_i^{(1)} - \tau_i \right| \leq w(N); \quad \max_{i=0,\ldots,J} |\hat{\nu}_i^{(1)} - \nu_i| \leq \rho_N \right\}, \tag{A.155}$$

Denote $G_N$ as the joint distribution of $J, \hat{\tau}_1^{(1)}, ..., \hat{\tau}_J^{(1)}, \hat{\nu}_0^{(1)}, ..., \hat{\nu}_J^{(1)}$; the domain of $G_N$ would be $\bigcup_{k=0}^{N-1} \mathbb{N}^{k+1} \times \mathbb{R}^{k+1}$.

Then, for any sequence $\{a_N\}$, we can bound $\mathbb{P}\left[\hat{J}=J; \quad \max_{i=1,...,J}\left|\hat{\tau}_i^{(2)}-\tau_i^{(2)}\right|\leq a_N\right]$ from below by:

$$
\begin{aligned}
& \mathbb{P}\left[\hat{J}=J; \quad \max_{i=1,...,J}\left|\hat{\tau}_i^{(2)}-\tau_i^{(2)}\right|\leq a_N\right] \\
\geq & \sum_{k=0}^{N-1} \int_{\substack{0<t_1<t_2<....<t_k<N \\ v_1,...,v_k\in\mathbb{R}}} \mathbb{P}\left[\hat{J}=J; \max_{i=1,...,J}\left|\hat{\tau}_i^{(2)}-\tau_i^{(2)}\right|\leq a_N \middle| \hat{J}=k, \hat{\tau}_j^{(1)}=t_j, \hat{\nu}_j^{(1)}=v_j \text{ for } j\leq k\right] \\
& dG_N(k,t_1,...,t_k,v_0,...,v_k) \\
\geq & \int_{\substack{|t_i-\tau_i|\leq Kw(N) \\ |v_i-\nu_i|\leq\rho_N \\ \text{for all } i}} \mathbb{P}\left[\hat{J}=J; \max_{i=1,...,J}\left|\hat{\tau}_i^{(2)}-\tau_i^{(2)}\right|\leq a_N \middle| \hat{J}=J, \hat{\tau}_j^{(1)}=t_j, \hat{\nu}_j^{(1)}=v_j \text{ for } j\leq J\right] \\
& dG_N(J,t_1,...,t_J,v_0,...,v_J) \\
\geq & \left(\inf_{\substack{|t_i-\tau_i|\leq Kw(N) \\ |v_i-\nu_i|\leq\rho_N \\ \text{for all } i}} \mathbb{P}\left[\hat{J}=J; \max_{i=1,...,J}\left|\hat{\tau}_i^{(2)}-\tau_i^{(2)}\right|\leq a_N \middle| \hat{J}=J, \hat{\tau}_j^{(1)}=t_j, \hat{\nu}_j^{(1)}=v_j \text{ for } j\leq J\right]\right) \cdot \\
& \mathbb{P}\left[\hat{J}=J; \quad \max_{i=1,...,J}\left|\hat{\tau}_i^{(1)}-\tau_i\right|\leq Kw(N); \quad \max_{i=0,...,J}|\hat{\nu}_i^{(1)}-\nu_i|\leq\rho_N\right] \\
\geq & \left(\inf_{\substack{|t_i-\tau_i|\leq Kw(N) \\ |v_i-\nu_i|\leq\rho_N \\ \text{for all } i}} \mathbb{P}\left[\max_{i=1,...,J}\left|\hat{\tau}_i^{(2)}-\tau_i^{(2)}\right|\leq a_N \middle| \hat{J}=J, \hat{\tau}_j^{(1)}=t_j, \hat{\nu}_j^{(1)}=v_j \text{ for } j\leq J\right]\right) - \\
& \qquad \mathbb{P}[\mathcal{R}_N \text{ is false}]
\end{aligned}
\tag{A.156}
$$

We wish to show that for all $\epsilon>0$, there exists a sequence $a_N=O(\log(J(N)))$ such that

$$\mathbb{P}\left[\hat{J}=J; \quad \max_{i=1,...,J}\left|\hat{\tau}_i^{(2)}-\tau_i^{(2)}\right|\leq a_N\right]>1-\epsilon$$

for all large $N$. It is sufficient to show this is satisfied by the second to last line of (A.156), as $\mathcal{R}_N$ is true with probability increasing to 1. Henceforth, we will work with the probability

$$\mathbb{P}\left[\max_{i=1,...,J}\left|\hat{\tau}_i^{(2)}-\tau_i^{(2)}\right|\leq a_N \middle| \hat{J}=J, \hat{\tau}_j^{(1)}=t_j, \hat{\nu}_j^{(1)}=v_j \text{ for } j\leq J\right]$$

and, in the domain $|t_i-\tau_i|\leq Kw(N)$ and $|v_i-\nu_i|\leq\rho_N$ for all $i$, we show that it is greater than $1-\epsilon$ for all sufficiently large $N$ and $a_N=C_1\log J+C_2$ for some

$C_1, C_2 > 0$. In the remainder of the proof, assume that all $t_i$'s and $v_i$'s fall within this domain.

For sufficiently large $N$, we have $Kw(N) \leq \delta_N/4$, and therefore no two of the second stage intervals ($[t_i - Kw(N),\, t_i + Kw(N)]$ for $i = 1, ..., J$) intersect. Because each $\hat{\tau}_j^{(2)}$ is a function of all $Y_i$'s in the disjoint index sets $S^{(2)}(t_j) \subset [t_j - Kw(N),\, t_j + Kw(N)]$ and the two level estimates $\hat{\nu}_{j-1}^{(1)}$ and $\hat{\nu}_j^{(1)}$, conditional independence holds:

$$\begin{aligned} & \mathbb{P}\left[ \max_{i=1,...,J} \left| \hat{\tau}_i^{(2)} - \tau_i^{(2)} \right| \leq a_N \middle| \hat{J} = J, \hat{\tau}_j^{(1)} = t_j, \hat{\nu}_j^{(1)} = v_j \text{ for } j \leq J \right] \\ = & \prod_{i=1}^{J} \mathbb{P}\left[ \left| \hat{\tau}_i^{(2)} - \tau_i^{(2)} \right| \leq a_N \middle| \hat{J} = J, \hat{\tau}_j^{(1)} = t_j, \hat{\nu}_j^{(1)} = v_j \text{ for all } j \right] \end{aligned}$$

To show the above product is eventually greater than some $1 - \varepsilon$, it would suffice to show that, for all $1 \leq k \leq J$ and sufficiently large $N$,

$$\begin{aligned} & \mathbb{P}_{J,\boldsymbol{v},\boldsymbol{t}}\left[ \left| \hat{\tau}_k^{(2)} - \tau_k^{(2)} \right| > a_N \right] \\ := & \mathbb{P}\left[ \left| \hat{\tau}_k^{(2)} - \tau_k^{(2)} \right| > a_N \middle| \hat{J} = J, \hat{\tau}_j^{(1)} = t_j, \hat{\nu}_j^{(1)} = v_j \text{ for all } j \right] \\ \leq & C_\epsilon / J \end{aligned} \tag{A.157}$$

for some $C_\epsilon < -\log(1 - \varepsilon)$.

For any $k$ between 1 to $J$ inclusive, we can write explicit expressions for $\tau_k^{(2)}$ and $\hat{\tau}_k^{(2)}$.

$$\begin{aligned}\hat{\tau}_k^{(2)} &= \underset{d\in S^{(2)}(t_k)}{\arg\min}\left(\operatorname{sgn}(v_k - v_{k-1})\sum_{i\in S^{(2)}(t_k)}\left(Y_i - \frac{v_{k-1}+v_k}{2}\right)\left[1(i\le d) - 1\left(i\le \tau_k^{(2)}\right)\right]\right)\\ &:= \underset{d\in S^{(2)}(t_k)}{\arg\min}\mathbb{M}_k(d) \qquad\qquad (A.158)\end{aligned}$$

Next, since, $t_k \in [\tau_k - w(N), \tau_k + w(N)]$, for $N$ large enough so that $\frac{K-1}{2}w(N) > 1$ we have

$$\begin{aligned}t_k - Kw(N) \le \tau_k - (K-1)w(N) < \tau_k^{(2)} - \frac{K-1}{2}w(N) <\\ \tau_k^{(2)} + \frac{K-1}{2}w(N) < \tau_k + (K-1)w(N) \le t_k + Kw(N) \qquad (A.159)\end{aligned}$$

That is to say, the set $S^{(2)}(t_k)$ includes the interval $\left[\tau_k^{(2)} - \frac{K-1}{2}w(N), \tau_k^{(2)} + \frac{K-1}{2}\right]$, minus the first stage subsample points, regardless of which $t_k$ was used among those permissible in $R_N$. Therefore, for any even integer $a_N$ such that $1 < a_N < \frac{K+1}{2}w(N)$ (which would be satisfied for all large $N$ if $a_N = O(\log(N)) = o(w(N))$),

$$\begin{aligned}&\mathbb{P}_{J,\boldsymbol{v},\boldsymbol{t}}\left[\left|\hat{\tau}_k^{(2)} - \tau_k^{(2)}\right| > a_N\right]\\ \le\ &\mathbb{P}_{J,\boldsymbol{v},\boldsymbol{t}}\left[\left|\lambda_2\left(\tau_k^{(2)}, \hat{\tau}_k^{(2)}\right)\right| > \frac{a_N}{2}\right] \qquad (A.160)\end{aligned}$$

for all sufficiently large $N$ such that the first stage subsample samples more sparsely than taking every other point. In order for $\left|\lambda_2\left(\hat{\tau}_k^{(2)}, \tau_k^{(2)}\right)\right| > a_N/2$, there must exists a $d$ such that

$$\left|\lambda_2(\tau_k^{(2)}, d)\right| > a_N/2 \qquad \text{and} \qquad \mathbb{M}_k(d) \le \min_{\substack{\ell\in S^{(2)}(t_k)\\ |\lambda(\tau_k^{(2)},\ell)|\le\frac{a_N}{2}}} \mathbb{M}_k(\ell) < \mathbb{M}_k(\tau_k^{(2)}) = 0 \qquad (A.161)$$

Therefore,

$$\begin{aligned}
& \mathbb{P}_{J,\boldsymbol{v},\boldsymbol{t}}\left[\left|\hat{\tau}_k^{(2)}-\tau_k^{(2)}\right|>a_N\right] \\
\leq\ & \mathbb{P}_{J,\boldsymbol{v},\boldsymbol{t}}\left[\left|\lambda_2\left(\tau_k^{(2)},\hat{\tau}_k^{(2)}\right)\right|>\frac{a_N}{2}\right] \\
\leq\ & \mathbb{P}_{J,\boldsymbol{v},\boldsymbol{t}}\left[\exists d\in S^{(2)}(t_k) \text{ where } \left|\lambda_2(\tau^{(2)},d)\right|>a_N/2 \text{ and } \mathbb{M}_k(d)<0\right] \\
\leq\ & \sum_{\substack{d\in S^{(2)}(t_k)\\ \left|\lambda_2(\tau^{(2)},d)\right|>a_N/2}} \mathbb{P}_{J,\boldsymbol{v},\boldsymbol{t}}\left[\mathbb{M}_k(d)\leq 0\right] \qquad \text{(A.162)}
\end{aligned}$$

Each $\mathbb{P}_{J,\boldsymbol{v},\boldsymbol{t}}\left[\mathbb{M}_k(d)\leq 0\right]$ can be bounded as follows by recognizing that each $\mathbb{M}_k(d)$ has subGaussian distribution: for every $d\in S^{(2)}(t_k)$ and sufficiently large $N$ so that $\rho_N\leq|\underline{\Delta}|/2$ (and hence $\mathrm{sgn}(v_{k-v_{k-1}})=\mathrm{sgn}(\nu_k-\nu_{k-1})$ for all $k$)

$$\begin{aligned}
& \mathbb{M}_k(d) \\
=\ & \mathrm{sgn}(v_k-v_{k-1})\sum_{i\in S^{(2)}(t_k)}\left(Y_i-\frac{v_{k-1}+v_k}{2}\right)\left[1(i\leq d)-1\left(i\leq\tau_k^{(2)}\right)\right] \\
=\ & \begin{cases}
\mathrm{sgn}(\nu_k-\nu_{k-1})\left(\lambda_2(\tau_k^{(2)},d)\left(\nu_k-\frac{v_{k-1}+v_k}{2}\right)+\sum\limits_{\substack{\tau_k^{(2)}<\ell\leq d\\ \ell\in S^{(2)}(t_k)}}\varepsilon_\ell\right) & \text{for } d>\tau_k^{(2)} \\
0 & \text{for } d=\tau_k^{(2)} \\
\mathrm{sgn}(\nu_k-\nu_{k-1})\left(\lambda_2(\tau_k^{(2)},d)\left(\nu_{k-1}-\frac{v_{k-1}+v_k}{2}\right)-\sum\limits_{\substack{d<\ell\leq\tau_k^{(2)}\\ \ell\in S^{(2)}(t_k)}}\varepsilon_\ell\right) & \text{for } d<\tau_k^{(2)}
\end{cases} \\
=\ & \begin{cases}
\left|\lambda_2(\tau_k^{(2)},d)\right|\left(\left|\frac{\nu_{k+1}-\nu_k}{2}\right|+\mathrm{sgn}(\nu_k-\nu_{k-1})\hat{D}_k\right)+\mathrm{sgn}(\nu_k-\nu_{k-1})\sum\limits_{\substack{\tau_k^{(2)}<\ell\leq d\\ \ell\in S^{(2)}(t_k)}}\varepsilon_\ell & \text{for } d>\tau_k^{(2)} \\
0 & \text{for } d=\tau_k^{(2)} \\
\left|\lambda_2(\tau_k^{(2)},d)\right|\left(\left|\frac{\nu_k-\nu_{k-1}}{2}\right|-\mathrm{sgn}(\nu_k-\nu_{k-1})\hat{D}_k\right)-\mathrm{sgn}(\nu_k-\nu_{k-1})\sum\limits_{\substack{d<\ell\leq\tau_k^{(2)}\\ \ell\in S^{(2)}(t_k)}}\varepsilon_\ell & \text{for } d<\tau_k^{(2)}
\end{cases}
\end{aligned} \qquad \text{(A.163)}$$

where

$$\hat{D}_k := \frac{(\nu_k - v_k) + (\nu_{k-1} - v_{k-1})}{2} \tag{A.164}$$

The maximal deviation of the signal estimates are at most $\rho_N \to 0$ away from the true signal estimates, and hence for all sufficiently large $N$, $\left|\hat{D}_k\right| \leq |\rho_N| \leq \frac{\underline{\Delta}}{2}$ (where $\underline{\Delta}$ bounds the minimum absolute jump from below). In this case

$$\left|\frac{\nu_{k+1} - \nu_k}{2} + \hat{D}_k\right| \geq \frac{\underline{\Delta}}{2} \quad \text{and} \quad \left|\frac{\nu_{k+1} - \nu_k}{2} - \hat{D}_k\right| \geq \frac{\underline{\Delta}}{2}. \tag{A.165}$$

Therefore the mean of the $\mathbb{M}_k(d)$'s satisfy

$$\mathbb{E}[\mathbb{M}_k(d)] = \left|\lambda_2(\tau_k^{(2)}, d)\right| \left(\left|\frac{\nu_{k+1} - \nu_k}{2}\right| \pm \operatorname{sgn}(\nu_k - \nu_{k-1})\hat{D}_k\right) \geq \left|\lambda_2(\tau_k^{(2)}, d)\right| \frac{\underline{\Delta}}{2}.$$

At the same time, $\mathbb{M}_k(d) - \mathbb{E}[\mathbb{M}_k(d)]$ is a zero mean subGaussian random variable with variance proxy parameter $\sigma^2 \left|\lambda_2(\tau_k^{(2)}, d)\right|$, hence

$$\begin{aligned}
& \mathbb{P}_{J,\boldsymbol{v},\boldsymbol{t}}\left[\mathbb{M}_k(d) \leq 0\right] \\
= \;& \mathbb{P}_{J,\boldsymbol{v},\boldsymbol{t}}\left[\mathbb{M}_k(d) - \mathbb{E}[\mathbb{M}_k(d)] \leq -\mathbb{E}[\mathbb{M}_k(d)]\right] \\
\leq \;& \mathbb{P}_{J,\boldsymbol{v},\boldsymbol{t}}\left[\mathbb{M}_k(d) - \mathbb{E}[\mathbb{M}_k(d)] \leq -\left|\lambda_2(\tau_k^{(2)}, d)\right| \frac{\underline{\Delta}}{2}\right] \\
\leq \;& \frac{1}{2}\exp\left(-\frac{\underline{\Delta}^2}{8\sigma^2}\left|\lambda_2(\tau_k^{(2)}, d)\right|\right)
\end{aligned} \tag{A.166}$$

Going back to (A.162), this gives:

$$\begin{aligned}
& \mathbb{P}_{J,\boldsymbol{v},\boldsymbol{t}}\left[\left|\hat{\tau}_k^{(2)}-\tau_k^{(2)}\right|>a_N\right] \\
\leq & \sum_{\substack{d\in S^{(2)}(t_k) \\ \left|\lambda_2(\tau^{(2)},d)\right|>a_N/2}} \mathbb{P}_{J,\boldsymbol{v},\boldsymbol{t}}\left[\mathbb{M}_k(d)\leq 0\right] \\
\leq & \frac{1}{2}\sum_{\substack{d\in S^{(2)}(t_k) \\ \left|\lambda_2(\tau^{(2)},d)\right|>a_N/2}} \exp\left(-\frac{\underline{\Delta}^2}{8\sigma^2}\left|\lambda_2(\tau_k^{(2)},d)\right|\right) \\
\leq & \frac{1}{2}\left[\sum_{k=\frac{a_N}{2}+1}^{\infty}\exp\left(-\frac{\underline{\Delta}^2}{8\sigma^2}k\right)+\sum_{k=-\frac{a_N}{2}-1}^{-\infty}\exp\left(-\frac{\underline{\Delta}^2}{8\sigma^2}|k|\right)\right] \\
= & \left(1-\exp\left(-\frac{\underline{\Delta}^2}{8\sigma^2}\right)\right)^{-1}\exp\left(-\frac{\underline{\Delta}^2}{8\sigma^2}\left(\frac{a_N}{2}+1\right)\right) \qquad \text{(A.167)}
\end{aligned}$$

Therefore in order to bound $\mathbb{P}_{J,\boldsymbol{v},\boldsymbol{t}}\left[\left|\hat{\tau}_k^{(2)}-\tau_k^{(2)}\right|>a_N\right]$ by $C_\varepsilon/J$ as we stated back in (A.157), a legitimate choice of $a_N$ which will satisfy this bound can be found by solving

$$\frac{\exp\left(-\frac{\underline{\Delta}^2}{8\sigma^2}\right)}{1-\exp\left(-\frac{\underline{\Delta}^2}{8\sigma^2}\right)}\exp\left(-\frac{\underline{\Delta}^2}{8\sigma^2}\cdot\frac{a_N}{2}\right)=\frac{C_\epsilon}{J}. \qquad \text{(A.168)}$$

The solution to this will be of the form $a_N = C_1\log(J)+C_2$, where $C_1$ and $C_2$ are constants not dependent on $N$. □

As one can see, this consistency theorem can be proven using only the tail probability properties of subGaussian variables. Another result that can be extended to such a setting would be Theorem II.5. Suppose the same conditions stated at the beginning of this section ((M1) to (M3), subGaussian errors, and (2.13)), condition (M5), and the condition that the error terms have a probability density with no point mass over $\mathbb{R}$ (i.e., $\mathbb{P}[\varepsilon_i = z] = 0\ \forall z\in\mathbb{R}$) , then the deviation of the final estimators will converge to distributions very similar to the $L_{\Delta/\sigma}$ distribution: for any integers

$k_1, \ldots, k_J$,

$$\mathbb{P}\left[\hat{J} = J;\ \lambda_2\left(\tau_j, \hat{\tau}_j^{(2)}\right) = k_j \text{ for } j = 1, \ldots, J\right] \to \prod_{j=1}^{J} \mathbb{P}\left[L_{\Delta_j}^* = k_j\right] \tag{A.169}$$

where $L_\Delta^* := \underset{t\in\mathbb{Z}}{\arg\min} X_\Delta^*(t)$, and $X_\Delta^*(t)$ is the random walk with absolute value drift

$$X_\Delta^*(t) := \begin{cases} \frac{t|\Delta|}{2} + \operatorname{sgn}(\Delta)\sum_{j=1}^{t}\varepsilon_j^* & t > 0 \\ 0 & t = 0 \\ \frac{|t\Delta|}{2} - \operatorname{sgn}(\Delta)\sum_{j=t+1}^{0}\varepsilon_j^* & t < 0 \end{cases} \tag{A.170}$$

where $\varepsilon_i^*$'s (for $i \in \mathbb{Z}$) have the same distribution as the error terms of the data sequence. The proof of this convergence in this subGaussian setting follows almost exactly as the proof found in section A.2.3. The only modification would be that $X_\Delta^*(t)$ and $L_\Delta^*$, defined in (A.115) and its following paragraph, would not necessarily equal the distribution of $X_{\Delta/\sigma}(t)$ and $L_{\Delta/\sigma}$, which were defined using Gaussian variables.

## A.3 Probability Bounds on Argmin of Random Walks Absolute Value Drifts

### A.3.1 Probability bound for Argmin of Random Walk

Here we will derive a probability bound for random walks of the form

$$X_\Delta(t) := \begin{cases} t\left|\frac{\Delta}{2}\right| + \sum_{i=1}^{t}\varepsilon_i & t > 0 \\ 0 & t = 0 \\ |t|\cdot\left|\frac{\Delta}{2}\right| - \sum_{i=-1}^{|t|}\varepsilon_i & t < 0 \end{cases} \tag{A.171}$$

Specifically, the following exponential bound applies:

**Lemma 5.** *Suppose that $S$ is a set of integers and $m$ a positive integer such that $[-m, m] \subset S$, then*

$$\mathbb{P}\left[\left|\arg\min_{t\in S} X_\Delta(t)\right| > m\right] \leq A(\Delta)\exp(-B(\Delta)m) \tag{A.172}$$

*where $A$ and $B$ are expressions dependent only on $|\Delta|$, with $A$ decreasing and $B$ increasing in $|\Delta|$.*

*Proof.*

$$\begin{aligned}
& \mathbb{P}\left[\left|\arg\min_{t\in S} X_\Delta(t)\right| > m\right] \\
= & \sum_{j>m,\, j\in S} \mathbb{P}\left[\arg\min_{t\in S} X_\Delta(t) = j\right] + \sum_{j<-m,\, j\in S} \mathbb{P}\left[\arg\min_{t\in S} X_\Delta(t) = j\right] \\
\leq & \sum_{t>m,\, t\in S} \mathbb{P}\left[X_\Delta(t) < 0\right] + \sum_{t<-m,\, t\in S} \mathbb{P}\left[X_\Delta(t) < 0\right] \\
\leq & \sum_{t>m,\, t\in S} \mathbb{P}\left[N\left(t\frac{|\Delta|}{2}, t\right) < 0\right] + \sum_{t<-m,\, t\in S} \mathbb{P}\left[N\left(|t|\frac{|\Delta|}{2}, |t|\right) < 0\right] \\
\leq & \sum_{t=m+1}^{\infty} \mathbb{P}\left[N(0,1) < -\frac{\sqrt{t}|\Delta|}{\sqrt{8}}\right] + \sum_{t=-m-1}^{-\infty} \mathbb{P}\left[N(0,1) < -\frac{\sqrt{|t|}|\Delta|}{\sqrt{8}}\right] \\
\leq & \sum_{t=m+1}^{\infty} \exp\left(-\frac{t\Delta^2}{8}\right) + \sum_{t=-m-1}^{-\infty} \exp\left(-\frac{t\Delta^2}{8}\right) \\
= & \ 2\left(\frac{\exp\left(-\frac{\Delta^2}{8}\right)}{1+\exp\left(-\frac{\Delta^2}{8}\right)}\right)\exp\left(-\frac{m\Delta^2}{8}\right)
\end{aligned} \tag{A.173}$$

□

This result has another implication. With probability approaching to 1 at an exponential pace, the argmin of $X_\Delta(t)$ equals the argmin over a smaller set:

**Lemma 6.** *For any set of integers $S$ and positive integer $m$ such that $[-m, m] \in S$,*

$$\mathbb{P}\left(\underset{t\in[-m,m]}{\arg\min} X_\Delta(t) = \underset{t\in S}{\arg\min} X_\Delta(t)\right) \geq 1 - A(\Delta)\exp(-B(\Delta)m) \tag{A.174}$$

*where $A$, $B$ are expressions in $\Delta$ that are, respectively, decreasing and increasing in $|\Delta|$ Lemma 5.*

*Proof.* The two argmins of $X_\Delta(t)$ (over $S$ and over $[m, m]$) are different if and only if the argmin over $S$ is outisde of the interval $[-m, m]$. Therefore

$$\begin{aligned} & \mathbb{P}\left(\underset{t\in[-m,m]}{\arg\min} X_\Delta(t) \neq \underset{t\in S}{\arg\min} X_\Delta(t)\right) \\ = \; & \mathbb{P}\left(\left|\underset{t\in S}{\arg\min} X_\Delta(t)\right| > m\right) \\ \leq \; & A(\Delta)\exp(-B(\Delta)m) \end{aligned} \tag{A.175}$$

□

This leads to

**Lemma 7.** *Suppose that $S$ is an integer set, $\ell$ and $m$ are positive integers, and $[-\ell, \ell] \subset [-m, m] \subset S$, then*

$$\left|\mathbb{P}\left[\left|\underset{t\in[-m,m]}{\arg\min} X_\Delta(t)\right| \leq \ell\right] - \mathbb{P}\left[\left|\underset{t\in S}{\arg\min} X_\Delta(t)\right| \leq \ell\right]\right| \leq A(\Delta)\exp(-B(\Delta)m) \tag{A.176}$$

*for some expressions $A()$ and $B()$ that are respectively, decreasing and increasing with respect to $|\Delta|$.*

*Proof.*

$$\begin{aligned} & \mathbb{P}\left[\left|\arg\min_{t\in[-m,m]} X_\Delta(t)\right| \le \ell\right] - \mathbb{P}\left[\left|\arg\min_{t\in S} X_\Delta(t)\right| \le \ell\right] \\ = \; & \mathbb{P}\left[\left|\arg\min_{t\in[-m,m]} X_\Delta(t)\right| \le \ell \text{ and } \left|\arg\min_{t\in S} X_\Delta(t)\right| > \ell\right] \\ \le \; & \mathbb{P}\left[\left|\arg\min_{t\in[-m,m]} X_\Delta(t)\right| \ne \left|\arg\min_{t\in S} X_\Delta(t)\right|\right] \end{aligned} \tag{A.177}$$

The last line is greater than 0, and by Lemma 6, less than $A(\Delta)\exp(-B(\Delta)m)$ for some appropriate expressions $A()$ and $B()$. □

### A.3.2 Quantiles

Due to Lemma 5, the following statement can be made regarding the quantile:

**Lemma 8.** *Using the A and B from Lemma 5,*

$$Q_\Delta(\sqrt[J]{1-\alpha}) \le \frac{1}{B}\log\frac{AJ}{\alpha} \tag{A.178}$$

*Proof.* Using the inequality from Lemma 5,

$$\begin{aligned} & \mathbb{P}\left[\left|\arg\min_{t\in\mathbb{Z}} X_\Delta(t)\right| \le \frac{1}{B}\log\frac{AJ}{\alpha}\right] \\ \ge \; & 1 - A\exp\left(-B\frac{1}{B}\log\frac{AJ}{\alpha}\right) \\ = \; & 1 - \frac{\alpha}{J} \\ \ge \; & \sqrt[J]{1-\alpha} \end{aligned} \tag{A.179}$$

□

### A.3.3 Comparison Between Random Walks, Part 1

Here will show some probability inequalities between random walks with different drifts. These results are useful in proving Theorem II.3.

**Lemma 9.** *Suppose that $S$ is a set of integers, $m$ is a positive integer, and $[-m, m] \subset S$. Define, for any positive $\Delta_1$, $\Delta_2$, the random walks*

$$W_{\Delta_1,\Delta_2}(t) = \begin{cases} |t|\frac{|\Delta_1|}{2} + \sum_{i=-1}^{t} \varepsilon_i & t < 0 \\ 0 & t = 0 \\ t\frac{|\Delta_2|}{2} + \sum_{i=1}^{t} \varepsilon_i & t > 0 \end{cases} \tag{A.180}$$

*Then for any $\eta > 0$,*

$$\begin{aligned} &\mathbb{P}\left[\left|\arg\min_{t\in S} W_{\Delta_1,\Delta_2}(t)\right| \leq m\right] \geq \mathbb{P}\left[\left|\arg\min_{t\in S} W_{\Delta_1,\Delta_2+2\eta}(t)\right| \leq m\right] \\ &- \left(A(\Delta_2)m^{3/2} + B(\Delta_2)\sqrt{m}\right)\eta\exp(-C(\Delta_2)m) \end{aligned} \tag{A.181}$$

*for some expressions $A(\Delta_2)$, $B(\Delta_2)$ and $C(\Delta_2)$ that are, respectively, decreasing, decreasing, and increasing with respect to $|\Delta_2|$. Similarly, the following inequality holds:*

$$\begin{aligned} &\mathbb{P}\left[\left|\arg\min_{t\in S} W_{\Delta_1,\Delta_2}(t)\right| \leq m\right] \geq \mathbb{P}\left[\left|\arg\min_{t\in S} W_{\Delta_1+2\eta,\Delta_2}(t)\right| \leq m\right] \\ &- \left(A(\Delta_1)m^{3/2} + B(\Delta_1)\sqrt{m}\right)\eta\exp(-C(\Delta_1)m) \end{aligned} \tag{A.182}$$

*where the form of the expressions $A()$, $B()$, and $C()$ has identical forms as expressions used in (A.181).*

*Proof.* It is only required to prove the inequality between $W_{\Delta_1,\Delta_2}$ and $W_{\Delta_1,\Delta_2+2\eta}$. This is because $W_{\Delta_1,\Delta_2}(t)$ has the same distribution as $W_{\Delta_2,\Delta_1}(-t)$ for all $t \in S$, and

therefore

$$\begin{aligned} & \mathbb{P}\left[\left|\arg\min_{t\in S} W_{\Delta_1,\Delta_2}(t)\right| \leq m\right] \\ = \; & \mathbb{P}\left[\left|\arg\min_{t\in S} W_{\Delta_2,\Delta_1}(-t)\right| \leq m\right] \\ = \; & \mathbb{P}\left[\left|\arg\min_{t\in -S} W_{\Delta_2,\Delta_1}(t)\right| \leq m\right] \\ & \text{where } -S := \{-t : t \in S\} \end{aligned} \tag{A.183}$$

where the last inequality is due to the fact that the existence of an $|\ell| \leq m$ such that $W_{\Delta_2,\Delta_1}(-\ell) < W_{\Delta_2,\Delta_1}(-t)$ for all $t \in S$, $|t| > m$ could be true if and only if there exists an $|\ell| \leq m$ such that $W_{\Delta_2,\Delta_1}(\ell) < W_{\Delta_2,\Delta_1}(t)$ for all $t \in -S$, $|t| > m$. Similarly, we also have

$$\mathbb{P}\left[\left|\arg\min_{t\in S} W_{\Delta_1+2\eta,\Delta_2}(t)\right| \leq m\right] = \mathbb{P}\left[\left|\arg\min_{t\in -S} W_{\Delta_2,\Delta_1+2\eta}(t)\right| \leq m\right], \tag{A.184}$$

and from here, an inequality can be derived by comparing

$$\mathbb{P}\left[\left|\arg\min_{t\in -S} W_{\Delta_2,\Delta_1}(t)\right| \leq m\right]$$

and

$$\mathbb{P}\left[\left|\arg\min_{t\in -S} W_{\Delta_2,\Delta_1+2\eta}(t)\right| \leq m\right]$$

using (A.181). Therefore, the rest of the proof will only concern the random walks $W_{\Delta_1,\Delta_2}(\cdot)$ and $W_{\Delta_1,\Delta_2+2\eta}(\cdot)$.

For the sake of brevity here, we will use the shorthand notations $W(t)$ for the random walk $W_{\Delta_1,\Delta_2}(t)$, and $W_+(t)$ for the random walk $W_{\Delta_1,\Delta_2+2\eta}(t)$. We are interested in the probability of the event when $\left|\arg\min_{t\in S} W(t)\right| > m$ and $\left|\arg\min_{t\in S} W_+(t)\right| \leq m$, so for now, assume that for some integer $k \in S$, such that $|k| > m$, $W(k) < W(t)$ for all

$t \in S$, $|t| \leq m$.

- If $k < -m$, then $W_+(k) = W(k) < W(t) \leq W(t) + t\eta 1(t > 0) = W_+(t)$ for all $t \in S - \{k\}$; in other words $\left|\arg\min_{t\in S} W_+(t)\right| > m$, a contradiction. Therefore it is not possible for $k < -m$.

- This leaves the possibility that $k > m$. Additionally:

  - as how '$k$' was defined, $W(k) < \min_{|t|\leq m} W(t)$

  - because $W_+(k) = W(k) + k\eta$ is not the minimum among the $W_+(t)$'s for $t \in S$, we have

$$\begin{aligned} W(k) + k\eta &= W_+(k) \\ &\geq \min_{|t|\leq m} W_+(t) \\ &\geq \min_{|t|\leq m} W(t) \end{aligned} \tag{A.185}$$

This breakdown of events shows that in order for the argmin of $W_+(t)$ to be within $[-m, m]$ but for the argmin of $W(t)$ to be outside this interval, there must be a $k > m$ where $\left(\min_{|t|\leq m} W(t)\right) - \eta k < W(k) \leq \min_{|t|\leq m} W(t)$. Therefore

$$\begin{aligned} & \mathbb{P}\left[\left|\arg\min_{t\in S} W(t)\right| > m \text{ and } \left|\arg\min_{t\in S} W_+(t)\right| \leq m\right] \\ \leq\ & \mathbb{P}\left[\exists k:\ k > m \text{ and } \left(\min_{|t|\leq m} W(t)\right) - \eta k < W(k) \leq \min_{|t|\leq m} W(t)\right] \\ \leq\ & \sum_{k\in S\cap(m,\infty)} \mathbb{P}\left[\left(\min_{|t|\leq m} W(t)\right) - \eta k < W(k) \leq \min_{|t|\leq m} W(t)\right]. \end{aligned} \tag{A.186}$$

The random variable $\min_{|t|\leq m} W(t)$ can either equal $\min_{t\in[0,m]} W(t)$ or $\min_{t\in[-m,0]} W(t)$. Therefore, for any specific $k$, the event $\left(\min_{|t|\leq m} W(t)\right) - \eta k < W(k) \leq \min_{|t|\leq m} W(t)$ implies that either $\left(\min_{t\in[0,m]} W(t)\right) - \eta k < W(k) \leq \min_{t\in[0,m]} W(t)$ or $\left(\min_{t\in[-m,0]} W(t)\right) -$

$\eta k < W(k) \le \min_{t\in[-m,0]} W(t)$, yielding the inequality

$$\begin{aligned}
& \mathbb{P}\left[\left(\min_{|t|\le m} W(t)\right) - \eta k < W(k) \le \min_{|t|\le m} W(t)\right] \\
\le\ & \mathbb{P}\left[\left(\min_{t\in[0,m]} W(t)\right) - \eta k < W(k) \le \min_{t\in[0,m]} W(t)\right] \\
& + \mathbb{P}\left[\left(\min_{t\in[-m,0]} W(t)\right) - \eta k < W(k) \le \min_{t\in[-m,0]} W(t)\right] \qquad \text{(A.187)}
\end{aligned}$$

Both of the two probabilities in the last part can be bounded. First, because $k > m$, $W(k)$ is independent of $W(-1), \ldots, W(-m)$, the distribution of $W(k)$ is still $N\left(k\frac{|\Delta_2|}{2}, k\right)$ even after conditioning on the value of $\min_{t\in[-m,0]} W(t)$, hence

$$\begin{aligned}
& \mathbb{P}\left[\left(\min_{t\in[-m,0]} W(t)\right) - \eta k < W(k) \le \min_{t\in[-m,0]} W(t)\right] \\
=\ & \mathbb{E}\left[\mathbb{P}\left[x - \eta k < W(k) \le x \middle| \min_{t\in[-m,0]} W(t) = x\right]\right] \\
=\ & \mathbb{E}\left[\mathbb{P}\left[\frac{x}{\sqrt{k}} - \sqrt{k}\left(\frac{|\Delta_2|}{2} + \eta\right) < N(0,1) \le \frac{x}{\sqrt{k}} - \sqrt{k}\frac{|\Delta_2|}{2} \middle| \min_{t\in[-m,0]} W(t) = x\right]\right] \\
=\ & \mathbb{E}\left[\int_{\frac{x}{\sqrt{k}} - \sqrt{k}\left(\frac{|\Delta_2|}{2}+\eta\right)}^{\frac{x}{\sqrt{k}} - \sqrt{k}\frac{|\Delta_2|}{2}} \frac{\exp(-z^2/2)}{\sqrt{2\pi}}\, dz \middle| \min_{t\in[-m,0]} W(t) = x\right] \\
\le\ & \int_{-\sqrt{k}\left(\frac{|\Delta_2|}{2}+\eta\right)}^{-\sqrt{k}\frac{|\Delta_2|}{2}} \frac{\exp(-z^2/2)}{\sqrt{2\pi}}\, dz \\
\le\ & \frac{\eta\sqrt{k}}{\sqrt{2\pi}} \exp\left[-\frac{\Delta_2^2}{8}k\right] \qquad \text{(A.188)}
\end{aligned}$$

As for the other inequality, consider the event that $\left(\min_{t\in[0,m]} W(t)\right) - \eta k < W(k) \le \min_{t\in[0,m]} W(t)$. Because $\min_{t\in[0,m]} W(t) \le W(0) = 0$, this event implies that for some $\ell \in [0, m]$, we have $W(\ell) - \eta k < W(k) < W(\ell) \le 0$ (namely, letting $\ell = \arg\min_{t\in[0,m]} W(t)$

would work). Therefore

$$\begin{aligned}
&\mathbb{P}\left[\left(\min_{t\in[0,m]} W(t)\right) - \eta k < W(k) \le \min_{t\in[0,m]} W(t)\right] \\
\le\ & \sum_{\ell=0}^{m} \mathbb{P}\left[W(\ell) - \eta k < W(k) \le W(\ell) \le 0\right] \\
=\ & \sum_{\ell=0}^{m} \mathbb{P}\left[-\eta k < W(k) - W(\ell) \le 0 \text{ and } W(\ell) \le 0\right] \qquad \text{(A.189)}
\end{aligned}$$

Now $W(\ell) = \ell\frac{|\Delta_2|}{2} + \sum_{j=1}^{\ell}\varepsilon_j$ and $W(k) - W(\ell) = (k-\ell)\frac{|\Delta_2|}{2} + \sum_{j=\ell+1}^{k}\varepsilon_j$ are independent random variables, with distributions $N\left(\ell\frac{|\Delta_2|}{2}, \ell\right)$ and $N\left((k-\ell)\frac{|\Delta_2|}{2}, (k-\ell)\right)$. Hence

$$\begin{aligned}
&\mathbb{P}\left[-\eta k < W(k) - W(\ell) \le 0 \text{ and } W(\ell) \le 0\right] \\
=\ & \mathbb{P}\left[-\eta k < W(k) - W(\ell) \le 0\right] \cdot \mathbb{P}\left[W(\ell) \le 0\right] \\
=\ & \mathbb{P}\left[-\sqrt{k-\ell}\frac{|\Delta_2|}{2} - \frac{\eta k}{\sqrt{k-\ell}} < N(0,1) \le -\sqrt{k-\ell}\frac{|\Delta_2|}{2}\right] \cdot \mathbb{P}\left[N(0,1) \le -\sqrt{\ell}\frac{|\Delta_2|}{2}\right] \\
=\ & \int_{-\sqrt{k-\ell}\frac{|\Delta_2|}{2} - \frac{\eta k}{\sqrt{k-\ell}}}^{-\sqrt{k-\ell}\frac{|\Delta_2|}{2}} \frac{\exp(-z^2/2)}{\sqrt{2\pi}}\, dz \cdot \mathbb{P}\left[N(0,1) \le -\sqrt{\ell}\frac{|\Delta_2|}{2}\right] \\
\le\ & \frac{\eta k}{\sqrt{k-\ell}} \frac{\exp\left(-\frac{\Delta_2^2}{8}(k-\ell)\right)}{\sqrt{2\pi}} \cdot \frac{1}{2}\exp\left[-\frac{\Delta_2^2}{8}\ell\right] \\
=\ & \frac{1}{2\sqrt{2\pi}} \cdot \frac{\eta k}{\sqrt{k-\ell}} \exp\left[-\frac{\Delta_2^2}{8}k\right] \qquad \text{(A.190)}
\end{aligned}$$

Therefore, using (A.189),

$$\begin{aligned}
& \mathbb{P}\left[\left(\min_{t\in[0,m]} W(t)\right) - \eta k < W(k) \le \min_{t\in[0,m]} W(t)\right] \\
\le\ & \sum_{\ell=0}^{m} \frac{1}{2\sqrt{2\pi}} \cdot \frac{\eta k}{\sqrt{k-\ell}} \exp\left[-\frac{\Delta_2^2}{8}k\right] \\
\le\ & \frac{\exp\left[-\frac{\Delta_2^2}{8}k\right]}{2\sqrt{2\pi}} \cdot \eta k \cdot \int_0^{m+1} \frac{1}{\sqrt{k-x}}\, dx \\
=\ & \frac{\exp\left[-\frac{\Delta_2^2}{8}k\right]}{\sqrt{2\pi}} \cdot \eta k \cdot (\sqrt{k} - \sqrt{k-m-1}) \\
\le\ & \frac{\exp\left[-\frac{\Delta_2^2}{8}k\right]}{\sqrt{2\pi}} \cdot \eta k \cdot \sqrt{m+1} \qquad \text{(A.191)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \mathbb{P}\left[\left|\arg\min_{t\in S} W(t)\right| > m \text{ and } \left|\arg\min_{t\in S} W_+(t)\right| \le m\right] \\
\le\ & \sum_{k\in S\cap(m,\infty)} \mathbb{P}\left[\left(\min_{t\in[0,m]} W(t)\right) - \eta k < W(k) \le \min_{t\in[0,m]} W(t)\right] \\
& + \sum_{k\in S\cap(m,\infty)} \mathbb{P}\left[\left(\min_{t\in[-m,0]} W(t)\right) - \eta k < W(k) \le \min_{t\in[-m,0]} W(t)\right] \\
& \text{according to (A.186) and (A.187)} \\
\le\ & \sum_{k=m+1}^{\infty} \left(\frac{\exp\left[-\frac{\Delta_2^2}{8}k\right]}{\sqrt{2\pi}} \cdot \eta k \cdot \sqrt{m+1}\right) + \sum_{k=m+1}^{\infty} \left(\frac{\eta\sqrt{k}}{\sqrt{2\pi}} \exp\left[-\frac{\Delta_2^2}{8}k\right]\right) \\
& \text{according to (A.188) and (A.191)} \\
\le\ & \frac{\eta(\sqrt{m+1}+1)}{\sqrt{2\pi}} \sum_{m+1}^{\infty} k \exp\left[-\frac{\Delta_2^2}{8}k\right] \\
\le\ & \frac{3\sqrt{m}}{\sqrt{2\pi}} \left(\frac{m\exp\left[-\frac{\Delta_2^2}{8}\right]}{1-\exp\left[-\frac{\Delta_2^2}{8}\right]} + \frac{\exp\left[-\frac{\Delta_2^2}{8}\right]}{\left(1-\exp\left[-\frac{\Delta_2^2}{8}\right]\right)^2}\right) \left(\eta \exp\left[-\frac{\Delta_2^2}{8}m\right]\right) \\
& \text{since } \sum_{k=a}^{\infty} k x^k = \left(\frac{a-1}{1-x} + \frac{1}{(1-x)^2}\right) x^a \text{ for any } |x|<1, a\in\mathbb{N} \qquad \text{(A.192)}
\end{aligned}$$

From here,

$$\begin{aligned}
& \mathbb{P}\left[\left|\arg\min_{t\in S} W(t)\right| \leq m\right] \\
\geq\ & \mathbb{P}\left[\left|\arg\min_{t\in S} W_+(t)\right| \leq m\right] - \mathbb{P}\left[\left|\arg\min_{t\in S} W_+(t)\right| \leq m \text{ and } \left|\arg\min_{t\in S} W(t)\right| > m\right] \\
\geq\ & \mathbb{P}\left[\left|\arg\min_{t\in S} W_+(t)\right| \leq m\right] - \left(A(\Delta_2)m^{3/2} + B(\Delta_2)\sqrt{m}\right)\eta\exp(-C(\Delta_2)m) \quad \text{(A.193)}
\end{aligned}$$

for some expressions $A(\Delta_2)$, $B(\Delta_2)$ and $C(\Delta_2)$ that are, respectively, decreasing, decreasing, and increasing with respect to $|\Delta_2|$. □

This result immediately leads to some results concerning the random walks $X_\Delta(\cdot)$, as they are a special case of the random walks $W_{\Delta_1,\Delta_2}(\cdot)$ where $\Delta_1 = \Delta_2$.

**Lemma 10.** *Suppose that for the random walk $Y_+(t)$ for $t \in S$ equals*

$$Y_+(t) = \begin{cases} X_\Delta(t) & \text{for } t \leq 0 \\ X_\Delta(t) + \eta t & \text{for } t > 0 \end{cases} \quad \text{(A.194)}$$

*for some constant $\eta$ such that $0 < \eta < \frac{|\Delta|}{2}$. Then for any $[-m, m] \subset S \subset \mathbb{Z}$,*

$$\begin{aligned}
\mathbb{P}\left[\left|\arg\min_{t\in S} Y_+(t)\right| \leq m\right] &\leq \mathbb{P}\left[\left|\arg\min_{t\in S} X_\Delta(t)\right| \leq m\right] \\
&+ \eta\left[A(\Delta)m^{3/2} + B(\Delta)\sqrt{m}\right]\exp\left[-C(\Delta)m\right] \quad \text{(A.195)}
\end{aligned}$$

*for some expressions $A()$, $B()$, and $C()$ which are, respectively, decreasing, decreasing, and increasing with respect to $|\Delta|$. The same probability inequality will hold if $Y_+(t) = X_\Delta(t) + \eta|t|1(t < 0)$.*

*A similar set of inequalities hold for the random walk $Y_-(t)$ for $t \in S$, defined as*

$$Y_-(t) = \begin{cases} X_\Delta(t) & \text{for } t \leq 0 \\ X_\Delta(t) - \eta t & \text{for } t > 0 \end{cases} \tag{A.196}$$

*For $\eta < \frac{|\Delta|}{2}$, and $[-m, m] \subset S$, we have*

$$\begin{aligned} &\mathbb{P}\left[\left|\arg\min_{t\in S} Y_-(t)\right| \leq m\right] \geq \mathbb{P}\left[\left|\arg\min_{t\in S} X_\Delta(t)\right| \leq m\right] - \\ &\eta\left[A\left(\frac{|\Delta|}{2} - \eta\right)m^{3/2} + B\left(\frac{|\Delta|}{2} - \eta\right)\sqrt{m}\right]\exp\left[-C\left(\frac{|\Delta|}{2} - \eta\right)m\right] \end{aligned} \tag{A.197}$$

*for some expressions $A()$, $B()$, and $C()$ which are, respectively, decreasing, decreasing, and increasing with respect to $|\Delta|/2 - \eta$. The same probability inequality will hold if $Y_-(t) = X_\Delta(t) - \eta|t|1(t < 0)$.*

*Proof.* Apply Lemma 9 with $\Delta_1 = \Delta_2 = \Delta$ to prove (A.195), and with $\Delta_1 = \Delta$, $\Delta_2 = \Delta - 2\eta$ to prove (A.197). □

Additionally, we can make probabilistic statements regarding the argmin of $X_\Delta(t)$ for two different values of $\Delta$:

**Lemma 11.** *For any $\Delta \neq 0$, $\eta > 0$, and a set $S$ which contains the interval $[-m, m]$,*

$$\mathbb{P}\left[\left|\arg\min_{t\in S} X_\Delta(t)\right| \leq m\right] \leq \mathbb{P}\left[\left|\arg\min_{t\in S} X_{|\Delta|+2\eta}(t)\right| \leq m\right] \tag{A.198}$$

*and*

$$\begin{aligned} &\mathbb{P}\left[\left|\arg\min_{t\in S} X_\Delta(t)\right| \geq m\right] \geq \mathbb{P}\left[\left|\arg\min_{t\in S} X_{|\Delta|+2\eta}(t)\right| \leq m\right] \\ &-2\eta\left[A(\Delta)m^{3/2} + B(\Delta)\sqrt{m}\right]\exp\left[-C(\Delta)m\right] \end{aligned} \tag{A.199}$$

*for some expressions $A()$, $B()$, and $C()$ which can take the same form and have the*

*same monotonicity properties as the ones used in Lemma 9.*

*Proof.* The first inequality can be shown by noticing that the event $\left|\arg\min_{t\in S} X_\Delta(t)\right| \leq m$ implies that for some $|\ell| \leq m$, $X_\Delta(\ell) \leq X_\Delta(t)$ for all $t \in S$, $|t| > m$. This in turn implies that $X_{|\Delta|+\eta}(\ell) = X_\Delta(\ell) + |\ell|\eta < X_\Delta(t) + |t|\eta = X_{|\Delta|+\eta}(t)$ for all $t \in S$, $|t| > m$, which means $\left|\arg\min_{t\in S} X_{|\Delta|+\eta}(t)\right| \leq m$.

The second inequality can be shown by applying Lemma 9 twice:

$$\begin{aligned}
& \mathbb{P}\left[\left|\arg\min_{t\in S} X_\Delta(t)\right| \leq m\right] \\
\geq\ & \mathbb{P}\left[\left|\arg\min_{t\in S} W_{|\Delta|,|\Delta|+2\eta}(t)\right| \leq m\right] - \eta\left[A(\Delta)m^{3/2} + B(\Delta)\sqrt{m}\right]\exp\left[-C(\Delta)m\right] \\
\geq\ & \mathbb{P}\left[\left|\arg\min_{t\in S} W_{|\Delta|+2\eta,|\Delta|+2\eta}(t)\right| \leq m\right] - 2\eta\left[A(\Delta)m^{3/2} + B(\Delta)\sqrt{m}\right]\exp\left[-C(\Delta)m\right] \\
=\ & \mathbb{P}\left[\left|\arg\min_{t\in S} X_{|\Delta|+\eta}(t)\right| \leq m\right] - 2\eta\left[A(\Delta)m^{3/2} + B(\Delta)\sqrt{m}\right]\exp\left[-C(\Delta)m\right] \quad \text{(A.200)}
\end{aligned}$$

□

### A.3.4 Comparison between Random Walks, Part 2

Here we will prove inequalities similar to those presented in Lemma 9, but in the other direction. These results are also useful in proving Theorem II.3.

**Lemma 12.** *Let the random walks $W_{\Delta_1,\Delta_2}$ be as they were defined in Lemma 9. Then given any positive $\eta$, positive integer $m$, and set $S$ such that $|\eta| < |\Delta_1|/2$ and $[-m, m] \subsetneq S$,*

$$\begin{aligned}
& \mathbb{P}\left[\left|\arg\min_{t\in S} W_{\Delta_1,\Delta_2+2\eta}(t)\right| \leq m\right] \geq \mathbb{P}\left[\left|\arg\min_{t\in S} W_{\Delta_1,\Delta_2}(t)\right| \leq m\right] \\
& + A\left(\frac{\Delta_1}{2} - \eta\right)\eta\sqrt{m}\exp\left(-B\left(\frac{\Delta_1}{2} - \eta\right)m\right) \quad \text{(A.201)}
\end{aligned}$$

*for some positive expressions $A()$ and $B()$ which are, respectively, decreasing and*

*increasing in* $\frac{\Delta_1}{2} - \eta$. *Similarly, between the random walks* $W_{\Delta_1,\Delta_2}$ *and* $W_{\Delta_1+2\eta,\Delta_2}$ *for* $0 < \eta < \frac{\Delta_2}{2}$ *there is the inequality*

$$\mathbb{P}\left[\left|\arg\min_{t\in S} W_{\Delta_1+2\eta,\Delta_2}(t)\right| \leq m\right] \geq \mathbb{P}\left[\left|\arg\min_{t\in S} W_{\Delta_1,\Delta_2}(t)\right| \leq m\right]$$
$$+A\left(\frac{\Delta_2}{2} - \eta\right)\eta\sqrt{m}\exp\left(-B\left(\frac{\Delta_2}{2} - \eta\right)m\right) \tag{A.202}$$

*for some positive expressions* $A()$ *and* $B()$ *which are, respectively, decreasing and increasing in* $\frac{\Delta_2}{2} - \eta$.

*Proof.* We will show the inequality between $W_{\Delta_1,\Delta_2+2\eta}$ and $W_{\Delta_1,\Delta_2}$, and the result between $W_{\Delta_1+2\eta,\Delta_2}$ and $W_{\Delta_1,\Delta_2}$ can be shown in a similar fashion or in an argument similar to what was found in the proof for Lemma 9. As in the proof of that lemma, we will use the shorthand notation $W$ for $W_{\Delta_1,\Delta_2}$ and $W_+$ for $W_{\Delta_1,\Delta_2+2\eta}$.

We are interested in how $\left|\arg\min_{t\in S} W(t)\right| \leq m$ and $\left|\arg\min_{t\in S} W_+(t)\right| > m$ can simultaneously occur, which we will do by considering the possible values of the argmin of $W(t)$. If these two events are true, then first note that for some integer $k \in [-m, m]$, $W(k) \leq W(t)$ for all $t \in S$, $t \neq k$.

- if $k \leq 0$, then $W_+(k) = W(k) \leq W(t) + |t|\eta 1(t > 0) = W_+(t)$ for all $t \in S$ and $|t| > m$; in other words $\left|\arg\min_{t\in S} W_+(t)\right| \leq m$,
- thus $k > 0$ and $W(k) = \arg\min_{t\in[0,m]} W(t)$. The only possible way for $\left|\arg\min_{t\in S} W_+(t)\right| > m$ is for the argmin to be less than $-m$: for any $t \in S$:
  - if $t > m$ then $W_+(t) = W(t) + t\eta > W(k) + k\eta = W_+(k)$, which means that the argmin of $W_+(t)$ cannot be greater than $m$ in absolute value
  - therefore the only possible argmin for $W_+(t)$ is for some $\ell < -m$, and it

must satisfy

$$W(\ell) = W_+(\ell) < W_+(k) = W(k) + \eta k \leq W(k) + \eta m \tag{A.203}$$

but at the same time since $W(\ell)$ was not the minimum among the $W(t)$'s, we have $W(\ell) \geq W(k)$

This breakdown of events shows that in order for the argmin of $W(t)$ to be within $[-m, m]$ but for the argmin of $W_+(t)$ to be outside this interval, there must be an $\ell < -m$ where $\min_{t\in[0,m]} W(t) < W(\ell) \leq \min_{t\in[0,m]} W(t) + \eta m$. Hence:

$$\begin{aligned}
&\mathbb{P}\left[\left|\arg\min_{t\in S} W(t)\right| \leq m \text{ and } \left|\arg\min_{t\in S} W_+(t)\right| > m\right] \\
&\leq \mathbb{P}\left[\exists \ell < -m \text{ where } \min_{t\in[0,m]} W(t) < W(\ell) \leq \min_{t\in[0,m]} W(t) + \eta m\right] \\
&\leq \sum_{\ell=-m-1}^{-\infty} \mathbb{P}\left[\min_{t\in[0,m]} W(t) < W(\ell) \leq \min_{t\in[0,m]} W(t) + \eta m\right] \\
&= \sum_{\ell=-m-1}^{-\infty} \mathbb{E}\left[\mathbb{P}\left[x < W(\ell) \leq x + \eta m \,\middle|\, \min_{t\in[0,m]} W(t) = x\right]\right] \\
&\leq \sum_{\ell=-m-1}^{-\infty} \mathbb{E}\left[\mathbb{P}\left[\frac{x}{\sqrt{|\ell|}} - \frac{\Delta_1}{2}\sqrt{|\ell|} < N(0,1) \leq \frac{x}{\sqrt{|\ell|}} - \frac{\Delta_1}{2}\sqrt{|\ell|} + \frac{\eta m}{\sqrt{|\ell|}} \,\middle|\, \min_{t\in[0,m]} W(t) = x\right]\right] \\
&= \sum_{\ell=-m-1}^{-\infty} \mathbb{E}\left[\int_{\frac{x}{\sqrt{|\ell|}} - \frac{\Delta_1}{2}\sqrt{|\ell|}}^{\frac{x}{\sqrt{|\ell|}} - \frac{\Delta_1}{2}\sqrt{|\ell|} + \frac{\eta m}{\sqrt{|\ell|}}} \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)\, dz \,\middle|\, \min_{t\in[0,m]} W(t) = x\right] \\
&\leq \sum_{\ell=-m-1}^{-\infty} \int_{-\frac{\Delta_1}{2}\sqrt{|\ell|}}^{-\frac{\Delta_1}{2}\sqrt{|\ell|} + \frac{\eta m}{\sqrt{|\ell|}}} \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)\, dz
\end{aligned}$$

since all possible values of $x$ are negative, and the integrated density is monotone

$$
\begin{aligned}
&\leq \sum_{\ell=-m-1}^{-\infty} \int_{-\frac{\Delta_1}{2}\sqrt{|\ell|}}^{-\frac{\Delta_1}{2}\sqrt{|\ell|}+\eta\sqrt{m}} \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)\, dz \\
&\leq \sum_{\ell=-m-1}^{-\infty} \frac{\eta}{\sqrt{2\pi}} \sqrt{m} \exp\left[-\frac{1}{2}\left(\frac{\Delta_1}{2}\sqrt{|\ell|} - \eta\sqrt{m}\right)^2\right] \\
&\leq \frac{\eta\sqrt{m}}{\sqrt{2\pi}} \sum_{\ell=-m-1}^{-\infty} \exp\left[-\frac{1}{2}\left(\frac{\Delta_1}{2} - \eta\right)^2 |\ell|\right] \\
&\leq \frac{1}{\sqrt{2\pi}} \left(\frac{\exp\left[-\frac{1}{2}\left(\frac{\Delta_1}{2} - \eta\right)^2\right]}{1 - \exp\left[-\frac{1}{2}\left(\frac{\Delta_1}{2} - \eta\right)^2\right]}\right) \eta\sqrt{m} \exp\left(-\frac{1}{2}\left(\frac{\Delta_1}{2} - \eta\right)^2 m\right).
\end{aligned}
\tag{A.204}
$$

Therefore,

$$
\begin{aligned}
& \mathbb{P}\left[\left|\arg\min_{t\in S} W_+(t)\right| \leq m\right] \\
\geq\ & \mathbb{P}\left[\left|\arg\min_{t\in S} W(t)\right| \leq m\right] - \mathbb{P}\left[\left|\arg\min_{t\in S} W(t)\right| \leq m \text{ and } \left|\arg\min_{t\in S} Y_+(t)\right| > m\right] \\
\geq\ & \mathbb{P}\left[\left|\arg\min_{t\in S} W(t)\right| \leq m\right] - A'\eta\sqrt{m}\exp(-B'm)
\end{aligned}
\tag{A.205}
$$

for some constants $A'$ and $B'$ depending only on $\frac{\Delta_1}{2} - \eta$.

□

We can immediately apply this result to random walks of the form $X_\Delta$:

**Lemma 13.** *Suppose that for the random walk $Y_+(t)$ for $t \in \mathbb{Z}$ equals*

$$
Y_+(t) = \begin{cases} X_\Delta(t) & \text{for } t \leq 0 \\ X_\Delta(t) + \eta t & \text{for } t > 0 \end{cases}
\tag{A.206}
$$

*for some constant* $\eta$ *such that* $0 < \eta < \frac{|\Delta|}{2}$. *Then for any* $[-m, m] \subset S \subset \mathbb{Z}$,

$$\begin{aligned}\mathbb{P}\left[\left|\arg\min_{t\in S} Y_+(t)\right| \leq m\right] &\geq \mathbb{P}\left[\left|\arg\min_{t\in S} X_\Delta(t)\right| \leq m\right] \\ &- A'\left(\frac{|\Delta|}{2} - \eta\right)\eta\sqrt{m}\exp\left(-B'\left(\frac{|\Delta|}{2} - \eta\right)m\right)\end{aligned} \tag{A.207}$$

*for some expressions* $A'()$ *and* $B'()$ *which are, respectively, decreasing and increasing with respect to* $\left(\frac{|\Delta|}{2} - \eta\right)$. *The same probability inequality will hold if* $Y_+(t) = X_\Delta(t) + \eta|t|1(t < 0)$.

*In addition, if* $Y_-(t)$ *is defined differently as*

$$Y_-(t) = \begin{cases} X_\Delta(t) & \text{for } t \leq 0 \\ X_\Delta(t) - \eta t & \text{for } t > 0 \end{cases} \tag{A.208}$$

*then we have the inequality*

$$\begin{aligned}\mathbb{P}\left[\left|\arg\min_{t\in S} Y_-(t)\right| \leq m\right] &\leq \mathbb{P}\left[\left|\arg\min_{t\in S} X_\Delta(t)\right| \leq m\right] \\ &+ A'\left(\frac{|\Delta|}{2} - \eta\right)\eta\sqrt{m}\exp\left(-B'\left(\frac{|\Delta|}{2} - \eta\right)m\right)\end{aligned} \tag{A.209}$$

*for some expressions* $A'()$ *and* $B'()$ *which are, respectively, decreasing and increasing with respect to* $\left(\frac{|\Delta|}{2} - \eta\right)$.

## A.4 Intelligent Sampling using Wild Binary Segmentation

### A.4.1 Wild Binary Segmentation

We next discuss the Wild Binary Segmentation (WBinSeg) algorithm, introduced in *Fryzlewicz et al.* (2014). Similar to our treatment of the BinSeg procedure, we will explain the WBinSeg procedure in the context of applying it to the dataset

$Z_1, \ldots, Z_{N^*}$, a size $\sim N^\gamma$ size subsample of a larger dataset $Y_1, \ldots, Y_N$ which satisfies the conditions (M1)-(M4). The steps of this algorithm are:

1. Fix a threshold value $\zeta_{N^*}$ and initialize the segment set $SS = \{(1, N)\}$, the change point estimate set $\hat{\underline{\tau}} = \emptyset$, and $M_{N^*}$ intervals $[s_1, e_1], \ldots, [s_{M_{N^*}}, e_{M_{N^*}}]$, where each $s_j$ and $e_j$ are uniformly picked from $\{1, \ldots, N^*\}$.

2. Pick any ordered pair $(s, e) \in SS$, remove it from $SS$ (update $SS$ by $SS \leftarrow SS - \{(s, e)\}$). If $s \geq e$ then skip to step 6, otherwise continue to step 3.

3. Define $\mathcal{M}_{s,e} := \{[s_i, e_i] : [s_i, e_i] \subseteq [s, e]\}$.

   - As an optional step, also take $\mathcal{M}_{s,e} \leftarrow \mathcal{M}_{s,e} \cup \{(s, e)\}$.

4. Find a $[s^*, e^*] \in \mathcal{M}_{s,e}$ such that

$$\max_{b \in \{s^*, \ldots, e^* - 1\}} |\bar{Y}^b_{s^*, e^*}| = \max_{[s', e'] \in \mathcal{M}_{s,e}} \left( \max_{b \in \{s', \ldots, e' - 1\}} |\bar{Y}^b_{s', e'}| \right)$$

   and let $b_0 = \underset{b \in \{s^*, \ldots, e^* - 1\}}{\arg\max} |\bar{Z}^b_{s^*, e^*}|$.

5. If $|\bar{Z}^{b_0}_{s^*, e^*}| \geq \zeta_{N^*}$, then add $b_0$ to the list of change point estimates (add $b_0$ to $\hat{\underline{\tau}}$), and add ordered pairs $(s, b_0)$ and $(b_0 + 1, e)$ to $SS$, otherwise skip to step 5.

6. Repeat steps 2-4 until $SS$ contains no elements.

Roughly speaking, WBinSeg performs very much like binary segmentation but with steps that maximize change point estimates over $M_{N^*}$ randomly chosen intervals. The consistency results in *Fryzlewicz et al.* (2014) imply that in our setting, the following holds:

**Theorem A.3.** *Suppose conditions (M1) to (M4) are satisfied and the tuning parameter $\zeta_{N^*}$ is chosen appropriately such that there exists positive constants $C_1$ and $C_2$*

*with* $C_1\sqrt{\log(N^*)} \leq \zeta_{N^*} \leq C_2\sqrt{\delta}_{N^*}$. *Denote* $\hat{J}$, $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{J}}$ *as the estimates obtained from wild binary segmentation. Then, there exists positive constants* $C_3, C_4$ *where*

$$\mathbb{P}\Big[\hat{J} = J; \quad \max_{i=1,\ldots,J} |\hat{\tau}_i - \tau_i| \leq C_3 \log(N^*)\Big] \geq 1 - C_4(N^*)^{-1} - \left(\frac{N^*}{\delta^*_{N^*}}\right)\left(1 - \left(\frac{\delta^*_{N^*}}{3N^*}\right)^2\right)^{M_{N^*}} \tag{A.210}$$

We remark that the right side of (A.210) does not necessarily converge to 1 unless $M_{N^*} \to \infty$ fast enough. Using some simple algebra, it was shown in the original paper, that for sufficiently large $N$, this expression can be bounded from below by $1 - CN^{-1}$ for some $C > 0$ if $M_{N^*} \geq \left(\frac{3N^*}{\delta_{N^*}}\right)^2 \log(N^{*2}/\delta_{N^*})$, a condition on $M_{N^*}$ which we we assume from here on in order to simplify some later analysis.

Compared with the consistency result for binary segmentation given in Theorem II.6, $\max_{j=1,\ldots,J} |\hat{\tau}_j - \tau_j|$ can be bounded by some constant times $\log(N^*)$, which can grow much slower than $E_{N^*} = (N^*/\delta^*_{N^*})^2 \log(N^*)$ whenever $N^*/\delta^*_{N^*} \to \infty$. However, this comes at the cost of computational time. Suppose we perform wild binary segmentation with $M_{N^*}$ random intervals $[s_1, e_1], \ldots, [s_{M_{N^*}}, e_{M_{N^*}}]$, then for large $N^*$ the most time consuming part of the operation is to maximize the CUSUM statistic over each interval, with the other tasks in the WBinSeg procedure taking much less time. This takes an order of $\sum_{j=1}^{M_{N^*}}(e_j - s_j)$ time, and since the interval endpoints are drawn from $\{1, \ldots, N^*\}$ with equal probability, we have $\mathbb{E}[e_j - s_j] = \frac{N^*}{3}(1 + o(1))$ for all $j = 1, \ldots, M_{N^*}$. Hence the scaling of the average computational time for maximizing the CUSUM statistic in all random intervals, and WBinSeg as a whole, is $O(N^* M_{N^*}) = O\left(\frac{N^{*3}}{(\delta^*_{N^*})^2} \log((N^*)^2/\delta^*_{N^*})\right)$ time, which is greater than $O(N^* \log(N^*))$ time for binary segmentation whenever $N^*/\delta^*_{N^*} \to \infty$. The trade-off between the increased accuracy of the estimates and the bigger computational time, as they pertain to intelligent sampling, will be analyzed later.

As with the BinSeg method, we need to verify that the WBinSeg method also satisfies the consistency condition of (2.13) at step (ISM3) of intelligent sampling, so that the results of Theorem II.3 continue to hold with the latter as the first stage procedure. To this end, we need to demonstrate a set of signal estimators that satisfy the condition of Lemma 4 with a $\rho^*_{N^*}$ such that $J\rho^*_{N^*} \to 0$. We do this by using the estimator proposed in (2.31), and also by imposing two further conditions:

(M8 (WBinSeg)): $\Xi$ (from condition (M3)) is further restricted by $\Xi \in [0, 1/3)$,

(M9 (WBinSeg)): $N_1$, from step (ISM2), is chosen so that $N_1 = K_1 N^\gamma$ for some $K_1 > 0$ and $\gamma > 3\Xi$.

**Lemma 14.** *Under conditions (M1) through (M4), (M8 (WBinSeg)), and (M9 (WBinSeg)), we have*

$$\mathbb{P}\left[\hat{J} = J; \quad \max_{i=1,\ldots,J} |\hat{\tau}_i - \tau_i| \leq w^*(N^*); \quad \max_{i=0,\ldots,J} |\hat{\nu}_i - \nu_i| \leq \rho^*_{N^*}\right] \to 1 \qquad \text{(A.211)}$$

*for some $\rho^*_{N^*}$ where $J\rho^*_{N^*} \to 0$.*

*Proof.* We need to show that wild binary segmentation does satisfy the requirements of Lemma 4. We have $\delta^*_{N^*} \gtrsim (N^*)^{1-\frac{\Xi}{\gamma}}$ and $J \lesssim (N^*)^{\frac{\Xi}{\gamma}}$. By the properties of the WBinSeg estimator shown in Theorem A.3, $w^*(N^*) \sim \log(N^*)$ and $B_{N^*} \sim N^{*-1}$. We shall show that for $\rho = (N^*)^\theta$ where $\theta \in \left(\frac{\Xi}{2\gamma} - \frac{1}{2}, -\frac{\Xi}{\gamma}\right)$, the conditions of Lemma 4 are satisfied, and in addition, $J\rho^*_{N^*} \to 0$.

First, the set $\left(\frac{\Xi}{2\gamma} - \frac{1}{2}, -\frac{\Xi}{\gamma}\right)$ is a valid set since

$$\frac{3\Xi}{2\gamma} - \frac{1}{2} = \frac{1}{2}\left(3\frac{\Xi}{\gamma} - 1\right) < 0. \qquad \text{(A.212)}$$

by condition (M9 (WBinSeg)). Hence $\frac{\Xi}{2\gamma} - \frac{1}{2} < -\frac{\Xi}{\gamma}$. Second, $J\rho^*_{N^*} \to 0$ since

$$J\rho^*_{N^*} \lesssim (N^*)^{\frac{\Xi}{\gamma}}(N^*)^{\theta} \tag{A.213}$$

But

$$\theta + \frac{\Xi}{\gamma} < 0.$$

Finally, as for the conditions of Lemma 4, we have

- $\frac{w^*(N^*)}{\delta^*_{N^*}} \lesssim (N^*)^{-(1-\frac{\Xi}{\gamma})} \log(N^*) \to 0$, hence $w^*(N^*) = o(\delta^*_{N^*})$
- $\frac{N^*}{\delta^*_{N^*}} B_{N^*} \sim \frac{1}{\delta^*_{N^*}} \to 0$
- because $\rho^*_{N^*} = (N^*)^{\theta}$ where $\theta \in \left(\frac{\Xi}{2\gamma} - \frac{1}{2}, -\frac{\Xi}{\gamma}\right)$, we have $\rho^*_{N^*} = o(1)$
- $\frac{N^* w^*(N^*)}{(\delta^*_{N^*})^{3/2}\rho^*_{N^*}} \lesssim (N^*)^{1-\theta}\log(N^*)$, and $\delta^*_{N^*}(\rho^*_{N^*})^2 \gtrsim (N^*)^{2\theta+1-\frac{\Xi}{\gamma}}$; since $2\theta+1-\frac{\Xi}{\gamma} > 0$, this means that $\frac{N^* w^*(N^*)}{(\delta^*_{N^*})^{3/2}\rho^*_{N^*}} = o(\exp(C_2\delta^*_{N^*}(\rho^*_{N^*})^2))$ for any positive constant $C_2$

Since all conditions of Lemma 4 are satisfied, the signal estimators satisfy

$$\mathbb{P}\left[\hat{J} = J;\ \max_{j=0,\ldots,J} |\hat{\nu}^*_j - \nu_j| \le \rho^*_{N^*}\right] \to 1 \tag{A.214}$$

which combines with the consistency of the change point estimators through a Bonferroni inequality to obtain the consistency result (2.13). □

**Remark 20.** *As with the BinSeg algorithm, the WBinSeg procedure is asymptotically consistent but faces the same issues as BinSeg in a practical setting where the goal is to obtain confidence bounds* $[\hat{\tau}_i \pm C_3\log(N^*)]$ *for the change point* $\tau_i$*. Namely, there are unspecified constants associated with the tuning parameter* $\zeta_{N^*}$ *and the confidence interval width* $C_3\log(N^*)$ *in (A.210). The issue of choosing a confidence interval width will can be resolved by applying the procedure of Section 2.4.2.*

Table A.1: Table of $\gamma_{min}$ and computational times for various values of $\Xi$, using WBinSeg at stage 1.

| $\Xi$ | $[0, 1/9)$ | $[1/9, 1/7)$ | $[1/7, 1/3)$ |
|---|---|---|---|
| $\gamma_{min}$ | $\frac{1-2\Xi+\Lambda}{2}$ | $\max\left\{\frac{1-2\Xi+\Lambda}{2}, 3\Xi+\eta\right\}$ | $\Xi+\eta$ |
| Order of Time | $N^{(1+2\Xi+\Lambda)/2}\log(N)$ | $N^{(1+2\Xi+\Lambda)/2}\log(N)$ or $N^{5\Xi+\eta}\log(N)$ | $N^{5\Xi+\eta}\log(N)$ |
| $\gamma_{min}$ $(\Lambda=0)$ | $\frac{1-2\Xi}{2}$ | $3\Xi+\eta$ | $3\Xi+\eta$ |
| Time $(\Lambda=0)$ | $N^{(1+2\Xi)/2}\log(N)$ | $N^{5\Xi+\eta}\log(N)$ | $N^{5\Xi+\eta}\log(N)$ |
| $\gamma_{min}$ $(\Lambda=\Xi)$ | $\frac{1-\Xi}{2}$ | $\frac{1-\Xi}{2}$ | $3\Xi+\eta$ |
| Time $(\Lambda=\Xi)$ | $N^{(1+3\Xi)/2}\log(N)$ | $N^{(1+3\Xi)/2}\log(N)$ | $N^{5\Xi+\eta}\log(N)$ |

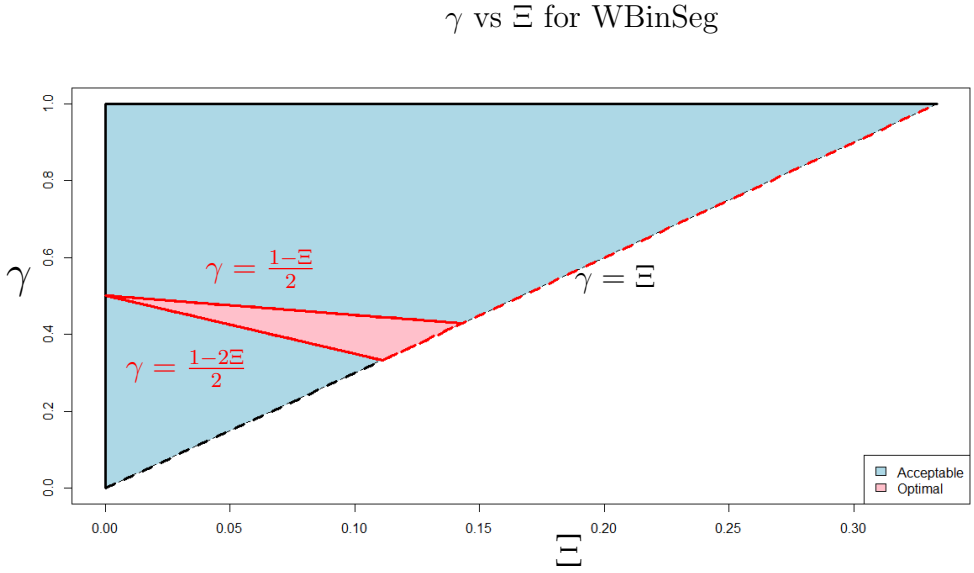


Figure A.3: Blue triangle encompasses all valid values of $\gamma$ vs $\Xi$ as set by (M8 (WBinSeg)). Pink region, solid red lines, and dotted red lines denotes $\gamma_{min}$ for each $\Xi$.

**Remark 21.** *Although the $\gamma_{min}$ values are, across the board, smaller than those given in Table 2.1 of the main chapter (which records the $\gamma_{min}$ values and computational time of Binseg at Stage 1), the order of the actual computational time is also greater across the board. There is some advantage in using WBinSeg since consistency condition (2.13) with $J\rho_N \to 0$ is satisfied in a greater regime of $\Xi$ ($\Xi < 1/3$ as opposed to $\Xi < 1/4$ for BinSeg), but in scenarios where the change points are placed far apart*

*BinSeg would run in shorter time.*

### A.4.2 Computational Time Order for Multiple Change Points

To analyze the computational time when using WBinSeg at stage 1, we again assume that $\delta_N/N^{\Xi} \to K_1$ and $J(N)/N^{\Lambda} \to K_2$ for some constant $\Lambda \leq \Xi$ and positive constants $K_1$, $K_2$. To summarize the details, for $N_1 \sim N^{\gamma}$, the average time for the first stage is $O(N^{\gamma+2\Xi}\log(N))$, while the second stage takes, on average, $O(N^{1-\gamma+\Lambda}\log(N))$ time. Together with condition (M9 (WBinSeg)) and setting, $\gamma > 3\Xi$, the order of average time for both stages combined is minimized by setting $\gamma_{min} = \max\left\{\frac{1-2\Xi+\Lambda}{2}, \Xi+\eta\right\}$ for any small constant $\eta$, with the average total computational time being $O(N^{\gamma_{min}+2\Xi}\log(N))$.

**Detailed Analysis:** We have $\delta_N \sim N^{1-\Xi}$ for some $\Xi \in [0, 1/3)$ and $J \sim N^{\Lambda}$ for some $\Lambda \in [0, \Xi]$. Given $n$ data points from (2.11) with minimum separation $\delta_n$ between change points, it takes an order of $\frac{n^3}{\delta_n^2}\log(n)$ time to perform the procedure, due to having to use $M_n \sim \left(\frac{n}{\delta_n}\right)^2 \log(n^2/\delta_n)$ random intervals, each requiring $O(n)$ time on average. The first stage works with a time series data of length order $N^{\gamma}$ with minimal separation $\delta^*_{N^*}$, and hence has an average computational time that is of the same order as$(N^*/\delta^*_{N^*})^2 \cdot N^*\log(N^*)$, which is the same order as $N^{\gamma+2\Xi}\log(N)$. The second stage works with $\hat{J}$ intervals, each of width $CN^{1-\gamma}\log(N)$ for some constant $C$. Because we have

$$\mathbb{P}[\hat{J} = J] \geq 1 - C(N^*)^{-1} \qquad \text{for some } C > 0 \tag{A.215}$$

by our earlier condition on $M_N$, we arrive at $\mathbb{E}[\hat{J}] = O(J)$ because

$$\mathbb{E}[\hat{J}] \leq J(1 - C(N^*)^{-1}) + N(C(N^*)^{-1}) \leq (C+1)J.$$

This in turn shows the expected computational time of the second stage is $O(JN^{1-\gamma}\log(N))$ which simplifies to $O(N^{1-\gamma+\Lambda}\log(N))$.

Both stages combined are expected to take $O\left(N^{(\gamma+2\Xi)\vee(1-\gamma+\Lambda)}\log(N)\right)$ time. This fact combined with the requirement $\gamma > 3\Xi$ lead to an optimal way to choose $\gamma$ to minimize the amount of computational time:

- On the region $\Xi < 1/9$ we can solve the equation $\gamma_{min} + 2\Xi = 1 - \gamma_{min} + \Lambda$ to get the minimizing $\gamma$ as $\gamma_{min} = \frac{1-2\Xi+\Lambda}{2}$, which satisfies $\gamma_{min} > 3\Xi$. This results in $O(N^{\frac{1+2\Xi+\Lambda}{2}}\log(N))$ computational time.
- On the region $\Xi \in [1/9, 1/7)$:
  - If $\frac{1-2\Xi+\Lambda}{2} > 3\Xi$, set $\gamma_{min} = \frac{1-2\Xi+\Lambda}{2}$ resulting in $O(N^{\frac{1+2\Xi+\Lambda}{2}}\log(N))$ computational time.
  - Otherwise if $\frac{1-2\Xi+\Lambda}{2} \leq 3\Xi$, set $\gamma_{min} = 3\Xi + \eta$, where $\eta > 0$ is small, for $O(N^{5\Xi+\eta}\log(N))$ computational time.
- For $\Xi \in [1/7, 1/3)$ also set $\gamma_{min} = \Xi+\eta$, where $\eta > 0$ is small, for $O(N^{5\Xi+\eta}\log(N))$ computational time.

### A.4.3 Simulation Results for WBinSeg

We next looked at how effective intelligent sampling with WBinSeg would work in practice, by running a set of simulations with the same set of model parameters as in Setup 2 of Section 2.7, and used the exact same method of estimation except with WBinSeg used in place of the Binseg algorithm. For the tuning parameters of WBinSeg, the same $\zeta_N$ was retained and the number of random intervals was taken as $M_n = 20000$. Although we could have used the theoretically prescribed value of $M_N = 9(N_1/\delta^*_{N_1})^2\log(N_1^2/\delta^*_{N_1})$, this turns out to be over 400,000 and is excessive as setting $M_N = 20,000$ gave accurate estimates.

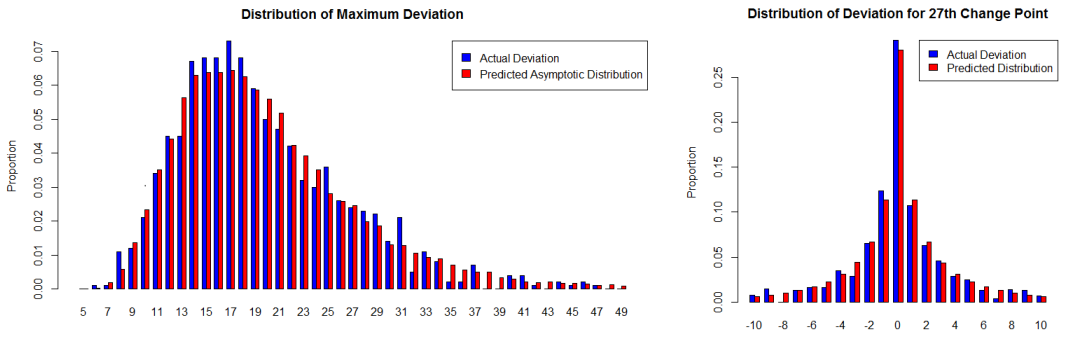


Figure A.4: Distributions of $\max_{1\leq j\leq 55}\lambda_2\left(\tau_j,\hat{\tau}_j^{(2)}\right)$ and $\lambda_2\left(\tau_{27},\hat{\tau}_{27}^{(2)}\right)$ from 1000 trials using the same parameters as setup 2 but employing WBinSeg instead of BinSeg.

Using WBinSeg along with steps (D1) and (D2), the event $\{\hat{J}=J\}$ also occurred over 99% of the time during simulations. One can also see from Figure A.4 that the distribution of the $\hat{\tau}_j^{(2)}$'s again match with Theorem II.5. Because the performance of intelligent sampling is near identical for this setup, regardless of whether BinSeg or WBinSeg was used, the reader may wonder why the latter isn't used for the previous few simulations. The reason is the following: when the re-fitting method from Section 2.4.2 is implemented, it results in the second stage intervals being of width $Q_1(1-\alpha/J)N^{1-\gamma}$, irrespective of which of BinSeg or WBinSeg was used at stage one [where $Q_1(1-\alpha/J)$ is the $1-\alpha/J$ quantile of $|L_1|$]. Hence, WBinSeg loses any possible advantage from the tighter confidence bound of width $O(\log(N))$ rather than $O(E_N)$ for BinSeg from stage one. So, in a sparse change point setting and with stage 1 refitting, WBinSeg provides no accuracy advantages but adds to the computational time, e.g., the 1000 iterations used to create Figure A.4 averaged $\approx 293$ seconds, while the iterations used to create Figure 2.11 averaged $\approx 7$ seconds.

# APPENDIX B

# Proofs for Section III

We first establish that the least squares expression is indeed minimized at the true parameters. In other words, if we denote the functions

$$f_{\alpha,\beta,\theta}(X,Y) := \big(Y - \alpha \cdot 1(X^T\theta \leq 0) - \beta \cdot 1(X^T\theta > 0)\big)^2 \tag{B.1}$$

for $0 < \alpha_0 < \beta_0 < 1$, $\theta_0 \in \mathbb{R}^d$, then $Pf_{\alpha,\beta,\theta}$ is minimized at $(\alpha_0, \beta_0, \theta_0)$.

**Lemma 15.** *For every $\alpha, \beta$, and $\theta$ inside the parameter space, $\mathbb{E}[f_{\alpha,\beta,\theta}(X,Y)]$ is minimized when $(\alpha,\beta,\theta) = (\alpha_0,\beta_0,\theta_0)$. Furthermore, if we define the distance function $d^* = d_n^*$ as*

$$\begin{aligned} d^*((\alpha_1,\beta_1,\theta_1),(\alpha_2,\beta_2,\theta_2)) &:= \sqrt{(\alpha_1-\alpha_2)^2 + (\beta_1-\beta_2)^2 + \|\theta_1-\theta_2\|_2}, \\ (\alpha_1,\beta_1,\theta_1),(\alpha_2,\beta_2,\theta_2) &\in \{(x,y) \in \mathbb{R}^2 : 0 < x < y < 1\} \times \mathbb{R}^p \end{aligned} \tag{B.2}$$

*then there exists a constant $K^-$, independent of $n$, such that*

$$\inf_{d^*((\alpha,\beta,\theta),(\alpha_0,\beta_0,\theta_0)) \geq \eta} (Pf_{\alpha,\beta,\theta} - Pf_{\alpha_0,\beta_0,\theta_0}) \geq K^-\eta^2 \tag{B.3}$$

*for all* $\eta \leq 1$. *There also exists a constant* $K^+$, *independent of* $n$, *such that*

$$\sup_{d^*((\alpha,\beta,\theta),(\alpha_0,\beta_0,\theta_0))\leq\eta} (Pf_{\alpha,\beta,\theta} - Pf_{\alpha_0,\beta_0,\theta_0}) \leq K^+\eta^2 \tag{B.4}$$

*for all* $\eta \leq 1$.

*Proof.* We start by noting that for all valid estimates $\alpha, \beta$, and $\theta$:

$$\begin{aligned}
& \mathbb{E}[f_{\alpha,\beta,\theta}(X,Y)] \\
= \;& \mathbb{E}\left[(Y-\alpha)^2 \cdot 1(X^T\theta \leq 0) + (Y-\beta)^2 \cdot 1(X^T\theta > 0)\right] \\
= \;& \mathbb{E}_X\left[\mathbb{E}\left[(Y-\alpha)^2 \cdot 1(X^T\theta \leq 0, X^T\theta_0 \leq 0)|X\right] + \mathbb{E}\left[(Y-\alpha)^2 \cdot 1(X^T\theta \leq 0, X^T\theta_0 > 0)|X\right]\right] \\
& + \mathbb{E}\left[\mathbb{E}\left[(Y-\beta)^2 \cdot 1(X^T\theta > 0, X^T\theta_0 \leq 0)|X\right] + \mathbb{E}\left[(Y-\beta)^2 \cdot 1(X^T\theta > 0, X^T\theta_0 > 0)|X\right]\right] \\
= \;& \left[(\alpha_0-\alpha)^2 + \alpha_0(1-\alpha_0)\right] P(X^T\theta \leq 0, X^T\theta_0 \leq 0) \\
& + \left[(\beta_0-\alpha)^2 + \beta_0(1-\beta_0)\right] P(X^T\theta \leq 0, X^T\theta_0 > 0) \\
& + \left[(\alpha_0-\beta)^2 + \alpha_0(1-\alpha_0)\right] P(X^T\theta > 0, X^T\theta_0 \leq 0) \\
& + \left[(\beta_0-\beta)^2 + \beta_0(1-\beta_0)\right] P(X^T\theta > 0, X^T\theta_0 > 0)
\end{aligned} \tag{B.5}$$

Therefore, suppose

$$\begin{aligned}
0 \;&=\; \mathbb{E}[f_{\alpha,\beta,\theta}(X,Y)] - \mathbb{E}[f_{\alpha_0,\beta_0,\theta_0}(X,Y)] \\
&=\; (\alpha_0-\alpha)^2 P(X^T\theta \leq 0, X^T\theta_0 \leq 0) + (\beta_0-\alpha)^2 P(X^T\theta \leq 0, X^T\theta_0 > 0) \\
&\quad + (\alpha_0-\beta)^2 P(X^T\theta > 0, X^T\theta_0 \leq 0) + (\beta_0-\beta)^2 P(X^T\theta > 0, X^T\theta_0 > 0)
\end{aligned} \tag{B.6}$$

All four terms on the last two lines must equal zero. This means that first

$$\begin{aligned}
0 = (\beta_0-\alpha)^2 P(X^T\theta \leq 0, X^T\theta_0 > 0) \geq (\beta_0-\alpha)^2 a^- \|\theta-\theta_0\|_2 \\
0 = (\alpha_0-\beta)^2 P(X^T\theta > 0, X^T\theta_0 \leq 0) \geq (\alpha_0-\beta)^2 a^- \|\theta-\theta_0\|_2
\end{aligned} \tag{B.7}$$

The only possible solution candidates are either $\theta = \theta_0$, or $(\alpha, \beta) = (\beta_0, \alpha_0)$ and

$\theta \neq \theta_0$. However, the latter is not a valid solution since $\alpha_0 < \beta_0$ and we are restricted to the domain where $\alpha < \beta$. Therefore $\theta = \theta_0$, and with the two remaining terms of (B.6) we are left with

$$\begin{aligned} 0 &= (\alpha_0 - \alpha)^2 P(X^T\theta_0 \leq 0) \\ 0 &= (\beta_0 - \beta)^2 P(X^T\theta_0 > 0) \end{aligned} \tag{B.8}$$

Both $P(X^T\theta_0 > 0)$ and $P(X^T\theta_0 \leq 0) = P(X^T(-\theta_0) \geq 0)$ must be strictly positive by our assumptions. Hence this means $\alpha = \alpha_0$ and $\beta = \beta_0$.

For the second assertion, suppose $d^*((\alpha, \beta, \theta), (\alpha_0, \beta_0, \theta_0)) \geq \eta$ for some $\eta \leq 1$, then from the above calculations

$$\begin{aligned} & Pf_{\alpha,\beta,\theta} - Pf_{\alpha_0,\beta_0,\theta_0} \\ = \; & (\alpha_0 - \alpha)^2 P(X^T\theta > 0, X^T\theta_0 > 0) + (\beta_0 - \alpha)^2 P(X^T\theta > 0, X^T\theta_0 \leq 0) \\ & + (\alpha_0 - \beta)^2 P(X^T\theta \leq 0, X^T\theta_0 > 0) + (\beta_0 - \beta)^2 P(X^T\theta \leq 0, X^T\theta_0 \leq 0) \end{aligned} \tag{B.9}$$

At least one of $(\alpha - \alpha_0)^2$, $(\beta - \beta_0)^2$, or $\|\theta - \theta_0\|_2$ must be greater than $\frac{\eta^2}{3}$. Suppose it is $\|\theta - \theta_0\|_2 > \frac{\eta^2}{3}$, then we must have

$$\begin{aligned} Pf_{\alpha,\beta,\theta} - Pf_{\alpha_0,\beta_0,\theta_0} &\geq (\alpha - \beta_0)^2 P(X^T\theta \leq 0, X^T\theta_0 > 0) + (\beta_0 - \alpha)^2 P(X^T\theta > 0, X^T\theta_0 \leq 0) \\ &\geq \left[(\alpha - \beta_0)^2 + (\beta_0 - \alpha)^2\right] \cdot a^- \|\theta - \theta_0\|_2 \\ &\geq \frac{a^-\left[(\alpha - \beta_0)^2 + (\beta_0 - \alpha)^2\right]}{3} \eta^2 \\ &\geq \frac{a^-(\beta_0 - \alpha_0)^2}{12} \eta^2 \end{aligned} \tag{B.10}$$

The last line can be deduced from

$$\inf_{0 \leq \alpha < \beta \leq 1} \left[(\alpha - \beta_0)^2 + (\beta_0 - \alpha)^2\right] \geq \frac{(\beta_0 - \alpha_0)^2}{4}$$

(this is because either $\beta > \frac{\alpha_0+\beta_0}{2}$ and $(\beta-\alpha_0)^2 > \frac{(\beta_0-\alpha_0)^2}{4}$, or $\beta \leq \frac{\alpha_0+\beta_0}{2}$, which means $\alpha < \frac{\alpha_0+\beta_0}{2}$ and hence $(\beta_0-\alpha)^2 > \frac{(\beta_0-\alpha_0)^2}{4}$). On the other hand, suppose $\|\theta-\theta_0\|_2 \leq \frac{\eta^2}{3}$, then we have

$$\|\theta+\theta_0\|_2 \geq 2\|\theta_0\|_2 - \|\theta-\theta_0\|_2 \geq 2 - \frac{\eta^2}{3} \geq 1 \tag{B.11}$$

and hence

$$\begin{aligned}
Pf_{\alpha,\beta,\theta} - Pf_{\alpha_0,\beta_0,\theta_0} &\geq (\alpha-\alpha_0)^2 P(X^T\theta > 0, X^T\theta_0 > 0) \\
&\quad +(\beta_0-\beta)^2 P(X^T\theta \leq 0, X^T\theta_0 \leq 0) \\
&= (\alpha-\alpha_0)^2 P(X^T(-\theta) < 0, X^T\theta_0 > 0) \\
&\quad +(\beta_0-\beta)^2 P(X^T(-\theta) \geq 0, X^T\theta_0 \leq 0) \\
&\geq (\alpha-\alpha_0)^2 \cdot a^- \|\theta+\theta_0\|_2 + (\beta_0-\beta)^2 \cdot a^- \|\theta+\theta_0\|_2 \\
&\geq a^-((\alpha-\alpha_0)^2 + (\beta_0-\beta)^2) \\
&\geq \frac{a^-}{3}\eta^2
\end{aligned} \tag{B.12}$$

For the final assertion, suppose $d^*((\alpha,\beta,\theta),(\alpha_0,\beta_0,\theta_0)) \leq \eta^2$ for some $\eta^2 \leq 1$, then

$$\begin{aligned}
& Pf_{\alpha,\beta,\theta} - Pf_{\alpha_0,\beta_0,\theta_0} \\
= \; & (\alpha_0-\alpha)^2 P(X^T\theta > 0, X^T\theta_0 > 0) + (\beta_0-\alpha)^2 P(X^T\theta > 0, X^T\theta_0 \leq 0) \\
& +(\alpha_0-\beta)^2 P(X^T\theta \leq 0, X^T\theta_0 > 0) + (\beta_0-\beta)^2 P(X^T\theta \leq 0, X^T\theta_0 \leq 0) \\
\leq \; & (\alpha_0-\alpha)^2 + P(X^T\theta > 0, X^T\theta_0 \leq 0) + (\beta_0-\beta)^2 + P(X^T\theta \leq 0, X^T\theta_0 > 0) \\
\leq \; & \eta^2 + 2a^+\|\theta-\theta_0\|_2 + \eta^2 \\
\leq \; & (2+2a^+)\eta^2
\end{aligned} \tag{B.13}$$

□

A secondary result, to be used later, results in a similar bound.

**Lemma 16.** *Suppose* $d^*((\alpha,\beta,\theta),(\alpha_0,\beta_0,\theta_0)) \leq \delta$ *for some* $\delta \in (0,1)$*, then*

$$P(f_{\alpha,\beta,\theta} - f_{\alpha_0,\beta_0,\theta_0})^2 \leq (2+2a^+)\delta^2 \tag{B.14}$$

*Proof.* From the derivations of Lemma 15, we have

$$\begin{aligned} & f_{\alpha,\beta,\theta} - f_{\alpha_0,\beta_0,\theta_0} \\ = \; & (\alpha-\alpha_0)^2 \mathbb{1}(X^T\theta \leq 0, X^T\theta_0 \leq 0) + (\beta_0-\alpha)\mathbb{1}(X^T\theta \leq 0, X^T\theta_0 > 0) \\ & + (\alpha_0-\beta)^2 \cdot \mathbb{1}(X^T\theta > 0, X^T\theta_0 \leq 0) + (\beta_0-\beta)^2 \cdot \mathbb{1}(X^T\theta > 0, X^T\theta_0 > 0). \end{aligned} \tag{B.15}$$

Therefore,

$$\begin{aligned} & (f_{\alpha,\beta,\theta} - f_{\alpha_0,\beta_0,\theta_0})^2 \\ = \; & (\alpha-\alpha_0)^4 \mathbb{1}(X^T\theta \leq 0, X^T\theta_0 \leq 0) + (\beta_0-\alpha)^2\mathbb{1}(X^T\theta \leq 0, X^T\theta_0 > 0) \\ & + (\alpha_0-\beta)^2 \cdot \mathbb{1}(X^T\theta > 0, X^T\theta_0 \leq 0) + (\beta_0-\beta)^4 \cdot \mathbb{1}(X^T\theta > 0, X^T\theta_0 > 0), \end{aligned} \tag{B.16}$$

which can yield the bound

$$\begin{aligned} & P(f_{\alpha,\beta,\theta} - f_{\alpha_0,\beta_0,\theta_0})^2 \\ \leq \; & (\alpha-\alpha_0)^4 + \mathbb{P}(X^T\theta \leq 0, X^T\theta_0 > 0) + \mathbb{P}(X^T\theta > 0, X^T\theta_0 \leq 0) + (\beta_0-\beta)^4 \\ \leq \; & 2\delta^4 + 2a^+\|\theta-\theta_0\|_2 \\ \leq \; & (2+2a^+)\delta^2 \end{aligned} \tag{B.17}$$

for $\delta \leq 1$. $\square$

### B.0.0.1 Proof of Theorem III.1

Consistency of the estimators will be proved by using bounds provided through a VC dimension argument. First we obtain a bound on the VC dimension of the set of $f_{\alpha,\beta,\theta}$ functions.

**Lemma 17.** *The class of functions* $\mathcal{F}_n := \{f_{\alpha,\beta,\theta} : 0 \leq \alpha < \beta \leq 1, \theta \in \mathbb{R}^p\}$*, with* $f_{\alpha,\beta,\theta}$ *as defined in (*B.*1), has VC dimension bounded by* $v_1 p + v_2$*, where* $v_1$ *and* $v_2$ *are positive universal constants.*

*Proof.* First we will show that the class of functions $\mathcal{C} := \{g_\alpha(y) = (y-\alpha)^2 : \alpha \in \mathbb{R}\}$ is a VC class with VC dimension 3. Suppose that for some $(y_1, t_1), (y_2, t_2), (y_3, t_3) \in \mathbb{R}^2$, $\mathcal{C}$ is able to pick out the sets $\{(y_2, t_2), (y_3, t_3)\}$, $\{(y_1, t_1), (y_3, t_3)\}$, and $\{(y_1, t_1), (y_2, t_2)\}$ with the subgraph sets $\{(y, t) : t < g_{\alpha_j}(y)\}$ for $j = 1, 2, 3$, respectively. Without loss of generalities we may assume that $\alpha_1 < \alpha_2 < \alpha_3$. We shall show that $\mathcal{C}$ cannot pick out the set $\{(t_2, y_2)\}$.

First, note that since $t_j \geq (y_j - \alpha_j)^2 \geq 0$ for $j = 1, 2, 3$, all $t_j$'s are non-negative. Second, we can order $y_1, y_2, y_3$, and $\alpha_2$ by noting that

$$\begin{aligned} & (y_1 - \alpha_1)^2 < t_1 \leq (y_1 - \alpha_2)^2 \\ \rightarrow \quad & |y_1 - \alpha_1| < |y_1 - \alpha_2| \\ \rightarrow \quad & y_1 < \frac{\alpha_1 + \alpha_2}{2} < \alpha_2 \end{aligned}$$

$$\begin{aligned} & (y_2 - \alpha_2)^2 < t_2 \leq (y_2 - \alpha_1) \\ \rightarrow \quad & y_2 > \frac{\alpha_1 + \alpha_2}{2} > y_1, \end{aligned} \tag{B.18}$$

then deducting that $y_2 < y_3$ and $\alpha_2 < y_3$ through a similar procedure. Now suppose

that there is an $\alpha \in \mathbb{R}$ which picks out the set $\{(t_2, y_2)\}$. Then we must have

$$\begin{aligned} & (y_3 - \alpha)^2 \leq t_3 < (y_3 - \alpha_2)^2 \\ \rightarrow \quad & |y_3 - \alpha_2| < y_3 - \alpha_2 \\ \rightarrow \quad & \alpha \in (\alpha_2, 2y_3 - \alpha_2) \end{aligned}$$

$$\begin{aligned} & (y_1 - \alpha)^2 \leq t_1 < (y_1 - \alpha_2)^2 \\ \rightarrow \quad & |y_1 - \alpha| < \alpha_2 - y_1 \\ \rightarrow \quad & \alpha \in (2y_1 - \alpha_2, \alpha_2), \end{aligned} \tag{B.19}$$

which is a contradiction on whether $\alpha$ is greater or less than $\alpha_2$. Hence we must deduce that $\mathcal{C}$ cannot pick out all subsets of $\{(y_1, t_1), (y_2, t_2), (y_3, t_3)\}$.

This means that the function class $\mathcal{C}_1 := \{g(x, y) = (y - \alpha) : 0 \leq \alpha < \beta \leq 1, \theta \in \mathbb{R}^p\}$ has VC dimension 3. We also know, from established results on VC classes, that the function class $\mathcal{C}_2 := \{g(x, y) = (|y| + 1)^2 \cdot 1(x^T\theta \leq 0) : 0 \leq \alpha < \beta \leq 1, \theta \in \mathbb{R}^p\}$ has VC dimension at most $p + 2$. This allows the conclusion that the function class

$$\mathcal{C}_3 = \{(y - \alpha)^2 \cdot 1(x^T\theta \leq 0) : 0 \leq \alpha < \beta \leq 1, \theta \in \mathbb{R}^p\} = \mathcal{C}_1 \wedge \mathcal{C}_2 \tag{B.20}$$

has VC dimension at most $p + 5$. A similar set of steps leads to the conclusion that

$$\mathcal{C}_4 := \{(y - \beta)^2 \cdot 1(x^T\theta > 0) : 0 \leq \alpha < \beta \leq 1, \theta \in \mathbb{R}^p\} \tag{B.21}$$

has VC dimension at most $p + 5$. Therefore $\mathcal{F}_n = \mathcal{C}_3 \vee \mathcal{C}_4$ has VC dimension at most $2p + 10$. □

Using this bound for the VC dimension, this allows us to show

**Lemma 18.** *$\|\mathbb{P} - P\|_{\mathcal{F}_n}$ converges to 0 in outer expectation (and thus also in probability).*

*Proof.* First note that $\mathcal{F}_n$ is bounded by the envelope $F_n(X, Y) := 1$. The class $\mathcal{F}_n$ is a VC class of functions, so using Theorem 2.6.7 of VDVW, we have

$$N(\epsilon \|F_n\|_{\mathbb{P}_n}, \mathcal{F}_n, L_1(\mathcal{P}_n)) \leq K_1 V(\mathcal{F}_n) \left(K_2/\epsilon\right)^{V(\mathcal{F}_n)} \tag{B.22}$$

where $V(\mathcal{F}_n)$ is the VC dimension of $\mathcal{F}_n$, and is no more than a constant multiple of $d$. Using the methods presented in Theorem 2.4.3 of VDVW,for any $\epsilon > 0$ we can find a $M_\epsilon$ such that $P^* F_n\{F_n > M_\epsilon\} \leq \epsilon$, and hence setting $\mathcal{F}_n^{M_\epsilon} := \{f\{|f| \leq M_\epsilon\} : f \in \mathcal{F}_n\}$, we have

$$\begin{aligned} \mathbb{E}^* \|\mathbb{P}_n - P\| &\leq 2\mathbb{E}_X \mathbb{E}_{\varepsilon^*} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i^* f(X_i) \right\|_{\mathcal{F}_n^{M_\epsilon}} + 2P^* F_n\{F_n > M_\epsilon\} \\ &\leq 2\mathbb{E}_X \sqrt{1 + \log N(\epsilon, \mathcal{F}_n^{M_\epsilon}, L_1(\mathbb{P}_n))} \sqrt{\frac{6}{n}} M_\epsilon + 3\epsilon \end{aligned} \tag{B.23}$$

where $\varepsilon_i^*$'s are iid Rademacher variables. The bracketing number $N(\epsilon, \mathcal{F}_n^{M_\epsilon}, L_1(\mathbb{P}_n))$ is no greater than the bracketing number $N(\epsilon, \mathcal{F}_n, L_1(\mathbb{P}_n))$, hence

$$\begin{aligned} & \sqrt{1 + \log N(\epsilon, \mathcal{F}_n^{M_\epsilon}, L_1(\mathbb{P}_n))} \\ \leq\ & \sqrt{1 + \log N(\epsilon, \mathcal{F}_n, L_1(\mathbb{P}_n))} \\ \leq\ & \sqrt{C_1 + C_2 \log(d) + C_3 d \log(\|F_n\|_{\mathbb{P}_n}/\epsilon)} \end{aligned} \tag{B.24}$$

for some $C_1, C_2, C_3$ not dependent on $n$. Because $x \to \sqrt{1 + \log(x)}$ is a concave function, the expression in (B.23) can be bounded by

$$2\sqrt{C_1 + C_2 \log(d) + C_3 d \log\left(\mathbb{E}_X \|F_n\|_{\mathbb{P}_n}/\epsilon\right)} \sqrt{\frac{6}{n}} M_\epsilon + 3\epsilon \tag{B.25}$$

through an application of Jensen's inequality. Since $d/n \to 0$, this expression goes to $3\epsilon$ as $n \to \infty$, and as $\epsilon$ was arbitrary, this means that $\mathbb{E}^* \|\mathbb{P}_n - P\|_{\mathcal{F}_n} \to 0$. □

Combining Lemmas 17 and 18 results in the consistency of the least squares estimators.

**Theorem B.1.** *Let $d^*$ be as defined as in (B.2). Then $d^*((\hat{\alpha}^{(sq)}, \hat{\beta}^{(sq)}, \hat{\theta}^{(sq)}), (\alpha_0, \beta_0, \theta_0)) \to 0$ in probability.*

*Proof.* From Lemma 15, there exists a constant $C$ not dependent on $n$ such that for any small $\epsilon > 0$,

$$\begin{aligned}
& d^*((\hat{\alpha}^{(sq)}, \hat{\beta}^{(sq)}, \hat{\theta}^{(sq)}), (\alpha_0, \beta_0, \theta_0)) \geq \epsilon \\
\to \quad & P f_{\hat{\alpha}^{(sq)}, \hat{\beta}^{(sq)}, \hat{\theta}^{(sq)}} - P f_{\alpha_0, \beta_0, \theta_0} \geq C\epsilon^2 \\
\to \quad & P f_{\alpha_0, \beta_0, \theta_0} - P f_{\hat{\alpha}^{(sq)}, \hat{\beta}^{(sq)}, \hat{\theta}^{(sq)}} - (\mathbb{P}_n f_{\alpha_0, \beta_0, \theta_0} - \mathbb{P}_n f_{\hat{\alpha}^{(sq)}, \hat{\beta}^{(sq)}, \hat{\theta}^{(sq)}}) \leq -C\epsilon^2 \\
\to \quad & \sup_{\substack{0 \leq \alpha < \beta \leq 1 \\ \|\theta\|_2 = 1}} |P f_{\alpha, \beta, \theta} - \mathbb{P}_n f_{\alpha, \beta, \theta}| \geq C\epsilon^2/2
\end{aligned} \tag{B.26}$$

Therefore,

$$\begin{aligned}
& P\left[d^*((\hat{\alpha}^{(sq)}, \hat{\beta}^{(sq)}, \hat{\theta}^{(sq)}), (\alpha_0, \beta_0, \theta_0)) \geq \epsilon\right] \\
\leq \quad & P\left[\sup_{\substack{0 \leq \alpha < \beta \leq 1 \\ \|\theta\|_2 = 1}} |P f_\alpha, \beta, \theta - \mathbb{P}_n f_\alpha, \beta, \theta| \geq C\epsilon^2/2\right] \\
\leq \quad & P\left[|P - \mathbb{P}_n|_{\mathcal{F}_n} \geq C\epsilon^2/2\right] \\
\to \quad & 0
\end{aligned} \tag{B.27}$$

□

Because $d^*((\hat{\alpha}^{(sq)}, \hat{\beta}^{(sq)}, \hat{\theta}^{(sq)}), (\alpha_0, \beta_0, \theta_0)) \xrightarrow{p} 0$ if and only if $\hat{\alpha}^{(sq)} \xrightarrow{p} \alpha_0$, $\hat{\beta}^{(sq)} \xrightarrow{p} \beta_0$, and $\|\hat{\theta}^{(sq)} - \theta_0\|_2 \xrightarrow{p} 0$, this demonstrates the results of Theorem III.1.

#### B.0.0.2 Rate Result Preliminaries

In this section as well as in several other sections, we will utilize a rate result which is an adaptation of a well-known result. A generic version can be found in *der Vaart and Wellner* (1996) as Theorem 3.2. We state the theorem here:

**Theorem B.2.** *Let $\mathbb{M}_n$ be a sequence of stochastic processes indexed by a sequence of semiparametric spaces $\Theta_n$ and $M_n : \Theta_n \to \mathbb{R}$ be a a sequence of deterministic functions, such that for every $n$ and for every $\tau$ in a neighborhood of $\tau_{0,n} \in \Theta_n$,*

$$M_n(\tau) - M_n(\tau_{0,n}) \lesssim -\rho_n^2(\tau, \tau_{0,n}) \tag{B.28}$$

*Suppose that for every $n$ and sufficiently small $\delta$,*

$$E^* \left[ \sup_{\rho_n(\tau,\tau_{0,n}) \leq \delta} |(\mathbb{M}_n - M_n)(\tau) - (\mathbb{M}_n - M_n)(\tau_{0,n})| \right] \lesssim \frac{\phi_n(\delta)}{\sqrt{n}} \tag{B.29}$$

*for functions $\phi_n$ such that $\delta \to \frac{\phi_n(\delta)}{\delta^\alpha}$ is decreasing for some $\alpha < 2$ not dependent on $n$. If $r_n$ is a sequence such that $r_n^2 \phi\left(\frac{1}{r_n}\right) \leq \sqrt{n}$, $\hat{\tau}_n$ is a sequence such $\mathbb{M}_{n,p_n^*}(\hat{\tau}_n) \geq \mathbb{M}_{n,p_n^*}(\theta_{\tau,n}) - O_p(r_n^{-2})$, and*

$$\mathbb{P}(\rho_n(\hat{\tau}_n, \tau_{0,n}) > \eta) \to 0 \tag{B.30}$$

*for any $\eta > 0$, then $r_n \rho_n(\hat{\tau}_n, \tau_{0,n}) = O_p(1)$.*

For the model under consideration, the notation of this theorem would correspond to

$$\begin{aligned} \mathbb{M}_n(\tau) &\leftarrow \mathbb{P}_n(f_{\alpha,\beta,\theta}(X,Y) - f_{\alpha_0,\beta_0,\theta_0}(X,Y)) \\ M_n(\tau) &\leftarrow P(f_{\alpha,\beta,\theta}(X,Y) - f_{\alpha_0,\beta_0,\theta_0}(X,Y)) \end{aligned} \tag{B.31}$$

In the least squares context, $(\hat{\alpha}^{(sq)}, \hat{\beta}^{(sq)}, \hat{\theta}^{(sq)})$ would be the minimizer of $\mathbb{M}_n$ and $(\alpha_0, \beta_0, \theta_0)$ would be the unique minimzer of $M_n$ (due to Lemma 15). Additionally let $\rho_n(\cdot, \cdot) = d^*(\cdot, \cdot)$, then

$$M_n(\alpha, \beta, \theta) - M_n(\alpha_0, \beta_0, \theta_0) > \rho_n^2((\alpha, \beta, \theta), (\alpha_0, \beta_0, \theta_0)) \tag{B.32}$$

for all $(\alpha, \beta, \theta)$ in a neighborhood of the true parameters. To utilize this theorem, a key ingredient is a bound for the left hand side of (B.29). Since the function class of interest, the collection of functions in (B.1), have a constant envelope function of 1 for all $n$, the following is sufficient in giving us this component:

**Theorem B.3.** *Let $\mathcal{F}$ be a measurable class of functions with a constant envelope $U$ such that for $A > e^2$ and $V \geq 2$ and for every finitely supported probability measure $Q$*

$$N(\epsilon U, \mathcal{F}, L_2(Q)) \leq \left(\frac{A}{\epsilon}\right)^V \quad 0 \leq \epsilon < 1$$

*Then, for all $n$,*

$$\mathbb{E}\left\|\sum_{i=1}^n (f(X_i) - Pf)\right\|_{\mathcal{F}} \leq L\left(\sigma\sqrt{nV\log\frac{AU}{\sigma}} \vee VU\log\frac{AU}{\sigma}\right)$$

*where $L$ is an universal constant and $\sigma$ is such that $\sup_{f\in\mathcal{F}} P(f - Pf)^2 \leq \sigma^2$. In particular if $n\sigma^2 \gtrsim V\log(AU/\sigma)$ then the above result shows that*

$$\mathbb{E}\left\|\sum_{i=1}^n (f(X_i) - Pf)\right\|_{\mathcal{F}} \lesssim \sqrt{n\sigma^2 V\log(AU/\sigma)}$$

#### B.0.0.3 Proof of Theorem III.2

Using the results of Section B.0.0.2, we are now ready to prove the rate of convergence of the least squares estimators shown in Theorem III.2, which was that $(\hat{\alpha}^{(sq)} - \alpha_0)^2$, $(\hat{\beta}^{(sq)} - \beta_0)^2$, and $\|\hat{\theta}^{(sq)} - \theta_0\|_2$ are $O_p\left(\frac{d}{n}\log\left(\frac{n}{d}\right)\right)$.

*Proof.* First, for all $\delta > 0$ lets temporarily denote the set of functions

$$\begin{aligned}\mathcal{F}_\delta &:= \{f_{\alpha,\beta,\theta} - f_{\alpha_0,\beta_0,\theta_0} : d^*((\alpha,\beta,\theta),(\alpha_0,\beta_0,\theta_0)) \leq \delta\} \\ \mathcal{G}_\delta &:= \{(f_{\alpha,\beta,\theta} - f_{\alpha_0,\beta_0,\theta_0})^2 : d^*((\alpha,\beta,\theta),(\alpha_0,\beta_0,\theta_0)) \leq \delta\}\end{aligned} \tag{B.33}$$

Using Theorem B.3 would give an upper bound for (B.29) so long as we can obtain an upper bound for

$$\sup_{f \in \mathcal{G}_\delta} Pf^2. \tag{B.34}$$

Now note that for every $(\alpha, \beta, \theta)$ that whose distance from the truth is less than $\delta$, we have

$$P(f_{\alpha,\beta,\theta} - f_{\alpha_0,\beta_0,\theta_0})^2 \leq C\delta^2 \tag{B.35}$$

for all $\delta$ sufficiently small, where $C$ is some constant.

Plugging in 1 for $U$, $C\delta$ for $\sigma^2$, using the VC bound in (B.22) would give

$$\mathbb{E}\frac{1}{\sqrt{n}}\left\|\sum_{i=1}^n (f(X_i) - Pf)\right\|_{\mathcal{F}_\delta} \lesssim \delta\sqrt{d\log\frac{A}{\delta}} \vee \frac{d}{\sqrt{n}}\log\frac{A}{\delta} \tag{B.36}$$

for some positive constant $A$. Therefore we have

$$\mathbb{E}^*\left[\sup_{\rho_n(\tau,\tau_{0,n})\leq\delta} |(\mathbb{M}_n - M_n)(\tau) - (\mathbb{M}_n - M_n)(\tau_{0,n})|\right] \lesssim \frac{\phi_{n,d_n}(\delta)}{\sqrt{n}} \tag{B.37}$$

where

$$\phi_n(\delta) : \delta\sqrt{d\log\frac{A}{\delta}} \vee \frac{d}{\sqrt{n}}\log\frac{A}{\delta} \tag{B.38}$$

and this indeed satisfies $\phi(\delta)/\delta^\alpha$ being a decreasing function for $\alpha \in (1,2)$. In order for

$$\begin{aligned} \sqrt{n} \geq r_n^2 \phi_n(r_n^{-1}) &= r_n^2 \left[ \left( \frac{1}{r_n} \sqrt{d \log(A r_n)} \right) \vee \left( \frac{d}{\sqrt{n}} \log(A r_n) \right) \right] \quad \text{or} \\ \sqrt{\frac{n}{d}} &\geq r_n \sqrt{\log(A r_n)} \quad \text{and} \quad \frac{n}{d} \geq \log(A r_n) \end{aligned} \tag{B.39}$$

which can be satisfied by the compound statement

$$\begin{aligned} r_n &\leq \sqrt{\frac{n}{p}} \left( \log \left( \frac{An}{d} \right) \right)^{-1/2} \\ r_n &\leq \frac{1}{A} \exp\left( \frac{n}{d} \right). \end{aligned} \tag{B.40}$$

The second line implies the first line, hence we can take $r_n$ to be a constant multiple of $\sqrt{\frac{n}{p}} \left( \log \left( \frac{An}{d} \right) \right)^{-1/2}$. Using Theorem B.2, we have

$$r_n d^*((\hat{\alpha}^{(sq)}, \hat{\beta}^{(sq)}, \hat{\theta}^{(sq)}), (\alpha_0, \beta_0, \theta_0)) = O_p(1), \tag{B.41}$$

which leads to the conclusion that

$$(\hat{\alpha}^{(sq)} - \alpha_0)^2, (\hat{\beta}^{(sq)} - \beta_0)^2, \|\hat{\theta}^{(sq)} - \theta_0\|_2 \text{ are } O_p\left( \frac{d}{n} \log\left( \frac{n}{d} \right) \right) \tag{B.42}$$

□

#### B.0.0.4 Proof of Theorem III.3

In the following note we assume that we know $\alpha_0, \beta_0$ and we want to estimate $\theta_0$. As $\theta_0$ is only identifiable upto its norm, we assume $\|\theta_0\| = 1$. We use Fano's inequality to prove that we can estimate at best at a rate $d/n$

**Theorem B.4** (Fano's inequality)**.** *If $M \subseteq \Theta$ is a finite $2\epsilon$ packing set, i.e. for any two $\theta_i, \theta_j \in \Theta'$, $\|\theta_i - \theta_j\|_2 \geq 2\epsilon$, then, based on $n$ i.i.d. samples $z_1, z_2, \ldots, z_n \sim P_\theta$ we*

*have the following minimax lower bound:*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}\left( \|\hat{\theta} - \theta\|^2 \right) \geq \epsilon^2 \left( 1 - \frac{\frac{n}{|M|^2} \sum_{i,j} KL(\mathbf{P}_{\theta_i} || \mathbf{P}_{\theta_j}) + \log 2}{\log\left(|M| - 1\right)} \right)$$

To use Fano's inequality we use the following set of alternatives:

$$M = \left\{ \frac{\theta}{\|\theta\|} : \theta = (1, \tilde{\theta}), \tilde{\theta} \in T \subset \mathbb{R}^{p-1} \right\}$$

where the set $T$ is a $2\epsilon$ packing set of a ball of radius $4\epsilon$ around origin in $\mathbb{R}^{d-1}$.

The following theorem gives us a hold on the number of elements of $M$:

**Theorem B.5.** *Following two properties are true for $M$:*

1. $|M| \geq 2^{d-1}$.
2. *For any $\theta_i, \theta_j \in M$ we have: $\epsilon/\sqrt{2} \leq \|\theta_i - \theta_j\| \leq 32\epsilon$.*

*Proof.* To prove the first part we resort to the following lemma which provides a lower bound on the packing number of a subset of $R^{d-1}$:

**Lemma 19.** *Let $\Theta \subset \mathbb{R}^{d-1}$. Define $D(\Theta, \|.\|_2, \epsilon)$ to be the $\epsilon$ packing set of $\Theta$ with respect to euclidean distance. Then we have: $D(\Theta, \|.\|_2, \epsilon) \geq \frac{1}{\epsilon^d} \frac{vol(\Theta)}{vol(B)}$, where $B$ is unit ball in $\mathbb{R}^d$.*

□

By the construction of $M$ we have used $T$ which is a $2\epsilon$ packing set of a ball of radius $4\epsilon$ around origin. Hence by previous lemma we have:

$$|T| = D(B_{4\epsilon}, \|\cdot\|, 2\epsilon) \geq \frac{1}{(2\epsilon)^{p-1}} \frac{\text{vol}(\Theta)}{\text{vol}(B)} = 2^{p-1}$$

The first part now follows from the fact that $|M| = |T|$.

To prove the second part, take $\theta_i = (1, \tilde{\theta}_i), \theta_j = (1, \tilde{\theta}_j)$. We already have $2\epsilon \leq \|\theta_i - \theta_j\| \leq 8\epsilon$. Now:

$$\begin{aligned}
\left\| \frac{\theta_i}{\|\theta_i\|} - \frac{\theta_j}{\|\theta_j\|} \right\|^2 &= \frac{\|\theta_i\|\theta_j\| - \theta_j\|\theta_i\|\|^2}{\|\theta_i\|^2\|\theta_j\|^2} \\
&\geq \frac{1}{4}\left[(\|\theta_i\| - \|\theta_j\|)^2 + \|\tilde{\theta}_i\|\theta_j\| - \tilde{\theta}_j\|\theta_i\|\|^2\right] \\
&\geq \frac{1}{4}\|\tilde{\theta}_i\|\theta_j\| - \tilde{\theta}_j\|\theta_i\|\|^2 \\
&= \frac{1}{4}\left\|\tilde{\theta}_i\left(\|\theta_j\| - \|\theta_i\|\right) + \|\theta_i\|\left(\tilde{\theta}_j - \tilde{\theta}_i\right)\right\|^2 \\
&= \frac{1}{4}\left[\|\tilde{\theta}_i\|^2\left(\|\theta_j\| - \|\theta_i\|\right)^2 + \|\theta_i\|^2\left\|\tilde{\theta}_j - \tilde{\theta}_i\right\|^2 + 2\left\langle\tilde{\theta}_i\left(\|\theta_j\| - \|\theta_i\|\right), \|\theta_i\|\left(\tilde{\theta}_j - \tilde{\theta}_i\right)\right\rangle\right] \\
&\geq \frac{1}{4}\left[4\epsilon^2 - 2\|\tilde{\theta}_i\|\,|\|\theta_j\| - \|\theta_i\||\,\|\theta_i\|\|\tilde{\theta}_j - \tilde{\theta}_i\|\right] \\
&\geq \frac{1}{4}\left[4\epsilon^2 - 2\|\theta_i\|\|\tilde{\theta}_i\|\|\tilde{\theta}_j - \tilde{\theta}_i\|^2\right] \\
&\geq \frac{1}{4}\left[4\epsilon^2 - 1024\epsilon^3\right] \\
&\geq \frac{\epsilon^2}{2} \qquad \text{[For small enough } \epsilon\text{]}
\end{aligned}$$

Hence upon normalization, the elements will form a packing set of $\epsilon/\sqrt{2}$.

The upper bound is relatively easy to calculate:

$$\begin{aligned}
\left\| \frac{\theta_i}{\|\theta_i\|} - \frac{\theta_j}{\|\theta_j\|} \right\| &= \frac{\|\theta_i\|\theta_j\| - \theta_j\|\theta_i\|\|}{\|\theta_i\|\|\theta_j\|} \\
&\leq \|\theta_i\|\theta_j\| - \theta_j\|\theta_i\|\| \\
&\leq \|\theta_i\left(\|\theta_j\| - \|\theta_i\|\right) + \|\theta_i\|\left(\theta_i - \theta_j\right)\| \\
&\leq \|\theta_i\|\,|\|\theta_j\| - \|\theta_i\|| + \|\theta_i\|\|\theta_i - \theta_j\| \\
&\leq 2\|\theta_i - \theta_j\|\|\theta_i\| \\
&\leq 32\epsilon
\end{aligned}$$

Now to control the Kullback-Liebler divergence we can invoke the following lemma:

**Lemma 20.** *If $P \sim Ber(p_1)$ and $Q \sim Ber(q_1)$ and if $\frac{1}{4} \leq p_1, q_1 \leq \frac{3}{4}$, then $KL(P||Q) \leq \frac{16}{3}(p_1 - q_1)^2$.*

One can immediately note from the lemma that $1/4, 3/4$ is nothing sacred. In fact if, $a_1 < p_1, q_1 < a_2$ then one can show that $KL(P||Q) \leq C(a_1, a_2)(p_1 - q_1)^2$ where $C(a_1, a_2)$ is some constant depending on $a_1, a_2$. For minimax construction, we assume $X \sim N(0, I_p)$ which obeys Assumption A1.

$$\begin{aligned} KL(\mathbf{P}_{\theta_I}||P_{\theta_J}) &= \mathbb{E}_X \left( KL(P_{\theta_I}(Y|X)||P_{\theta_J}(Y|X)) \right) \\ &\leq C(\alpha_0, \beta_0)\mathbb{E}_X(P_{\theta_I}(Y|X) - P_{\theta_J}(Y|X))^2 \\ &= C(\alpha_0, \beta_0)(\alpha_0 - \beta_0)^2 E_X \left( \mathbb{1}_{\text{sgn}(X'\theta_I) \neq \text{sgn}(X'\theta_J)} \right) \\ &= C(\alpha_0, \beta_0)(\alpha_0 - \beta_0)^2 \mathbf{P} \left( \text{sgn}(X'\theta_I) \neq \text{sgn}(X'\theta_J) \right) \\ &\leq a_+ C(\alpha_0, \beta_0)(\alpha_0 - \beta_0)^2 \|\theta_I - \theta_J\|_2 \\ &\leq 8 a_+ C(\alpha_0, \beta_0)(\alpha_0 - \beta_0)^2 \epsilon \end{aligned} \tag{B.43}$$

Putting this in Theorem 1, we get,

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E} \left( \|\hat{\theta} - \theta\|^2 \right) \gtrsim \epsilon^2 \left( 1 - \frac{n\epsilon + \log 2}{\log \left( 2^{d-1} - 1 \right)} \right) \approx \epsilon^2 \left( 1 - \frac{n\epsilon}{d} \right)$$

Hence a valid choice for $\epsilon = p/2n$ wihch makes

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E} \left( \|\hat{\theta} - \theta\|^2 \right) \gtrsim \left( \frac{d}{n} \right)^2$$

### B.0.1 Proofs for Section 3.4

We next prove the results presented in Section 3.4. Before we focus on any specific theorem, we first validate a claim made at the beginning of the section, namely that under Assumption A1, and for any $\gamma \in (\alpha_0, \beta_0)$,

$$\mathbb{E}(Y - \gamma)\mathbb{1}(X^T\theta \leq 0) \tag{B.44}$$

is uniquely minimized at $\theta = \theta_0$. To see this, simply expand by taking conditional expectations:

$$\begin{aligned} & E[(Y - \gamma \cdot 1(X^T\theta \leq 0))] \\ = \; & E\left[E\left[(Y-\gamma)\cdot 1(X^T\theta \leq 0, X^T\theta_0 \leq 0)\middle|X\right]\right] \\ & + E\left[E\left[(Y-\gamma)\cdot 1(X^T\theta \leq 0, X^T\theta_0 > 0)\middle|X\right]\right] \\ = \; & E\left[(\alpha_0(X)-\gamma)\cdot 1(X^T\theta \leq 0, X^T\theta_0 \leq 0)\right] + E\left[(\beta_0(X)-\gamma)\cdot 1(X^T\theta \leq 0, X^T\theta_0 > 0)\right] \end{aligned} \tag{B.45}$$

and hence

$$\begin{aligned} & E[(Y - \gamma \cdot 1(X^T\theta \leq 0))] - E[(Y - \gamma \cdot 1(X^T\theta_0 \leq 0))] \\ = \; & E\left[(\beta_0(X)-\gamma)\cdot 1(X^T\theta \leq 0, X^T\theta_0 > 0)\right] - E\left[(\alpha_0(X)-\gamma)\cdot 1(X^T\theta > 0, X^T\theta_0 \leq 0)\right] \\ \geq \; & 0. \end{aligned} \tag{B.46}$$

This last inequality is strict, stemming from the fact $P(X^T\theta > 0, X^T\theta_0 \leq 0) > 0$ or $P(X^T\theta \leq 0, X^T\theta_0 > 0) > 0$ for all $\theta$ (due to Assumption A1).

#### B.0.1.1 Proof of Theorem III.4

In this section we will show that $\hat{\theta}$ is consistent. First, denote $\gamma_{mean}$ as the expectation of $\mathbb{E}[Y]$, or

$$\gamma_{mean} = \mathbb{E}Y = \alpha_0 \mathbb{P}[X^T\theta_0 \leq 0] + \beta_0 \mathbb{P}[X^T\theta_0 > 0]. \tag{B.47}$$

Implicitly, the value of $\gamma_{mean}$ could change with the value of $n$. Using this notation, we next denote the risk function

$$R_d(\theta) = \mathbb{E}((Y - \gamma_{mean})\mathbb{1}(X^T\theta \leq 0))$$

where $d$ is the dimension of $X$ (and $\theta$ as well), and the following processes:

1. $\mathbb{M}_{n,d}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})\mathbb{1}(X_i^T\theta \leq 0)$
2. $L_{n,d}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \gamma_{mean})\mathbb{1}(X_i^T\theta \leq 0)$
3. $B_{n,d}(\theta) = (\bar{Y} - \gamma_{mean})\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_i^T\theta \leq 0)$

The estimator would then satisfy

$$\hat{\theta} = \underset{\theta \in S^{d-1}}{\arg\max}\mathbb{M}_{n,d}(\theta)$$

Proving the consistency of $\hat{\theta}$ proceeds in the same way as proving the consistency of $\hat{\theta}^{(sq)}$, by first bounding the maximum difference of $\mathbb{M}_{n,d}$ and $R_d$ and the difference of $R_d(\theta)$ from $R_d(\theta_0)$, and then combining the two results to show that $\|\hat{\theta} - \theta_0\|_2$ converges to 0 in probability. We proceed with the first part of this method.

**Lemma 21.** $\sup_{\theta \in S^{d-1}} |\mathbb{M}_{n,d}(\theta) - R_d(\theta)| \xrightarrow{P} 0$ *as* $n \to \infty$.

*Proof.*

$$\begin{aligned}
\sup_{\theta \in S^{d-1}} |\mathbb{M}_{n,d}(\theta) - R_d(\theta)| &= \sup_{\theta \in S^{d-1}} |L_{n,d}(\theta) + B_{n,d}(\theta) - R_d(\theta)| \\
&\leq \sup_{\theta \in S^{d-1}} |L_{n,d}(\theta) - R_d(\theta)| + \sup_{\theta \in S^{d-1}} |B_{n,d}(\theta)| \\
&= T_1 + T_2
\end{aligned}$$

Next we analyze each summand separately. Define $\mathcal{F}_\theta = \{f_\theta : \theta \in S^{d-1}, f_\theta(x, y) = (y - \gamma_{mean})\mathbb{1}(x^T\theta \leq 0)\}$. Each $\mathcal{F}_\theta$ has a VC dimension bounded by a (universal) constant times $d$, and therefore we have

$$\begin{aligned}
T_1 &= \sup_{\theta \in S^{d-1}} |L_{n,d}(\theta) - R_d(\theta)| \\
&= \|P_n - P\|_{\mathcal{F}_\theta} \xrightarrow{P} 0 \quad \text{since } \frac{d}{n} \to 0
\end{aligned}$$

On the other hand,

$$\begin{aligned} T_2 &= \sup_{\theta \in S^{d-1}} |B_{n,d}(\theta)| \\ &= |\bar{Y} - \gamma_{mean}| \sup_{\theta \in S^{d-1}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i^T \theta \leq 0) \right| \\ &\leq |\bar{Y} - \gamma_{mean}| \xrightarrow{P} 0 \text{ as } n \to \infty. \end{aligned}$$

The last line is valid because even if $\gamma_{mean}$ changes with $n$, the variance of $\bar{Y} - \gamma_{mean}$, which is $\frac{\gamma_{mean}(1-\gamma_{mean})}{n}$, is always bounded above by $\frac{1}{4n}$. Even if the distribution of $Y$ changes with $n$, the variance of $(\bar{Y} - \gamma_{mean})$ nevertheless goes to 0. □

**Lemma 22.** *The curvature of the risk function satisfies the following constraint:*

$$\inf_{\|\theta - \theta_0\|_2 \geq \epsilon} [R(\theta) - R(\theta_0)] \geq \epsilon \frac{(\beta_0 - \alpha_0)a^-}{2}$$

*where $a^-$ comes from our Assumption A1.*

*Proof.*

$$\begin{aligned} R(\theta) - R(\theta_0) &= \mathbb{E}\left((Y - \gamma_{mean})\left(\mathbb{1}(X^T\theta \leq 0) - \mathbb{1}(X^T\theta_0 \leq 0)\right)\right) \\ &= \mathbb{E}_X\big((\alpha_0 - \gamma_{mean})\left[\mathbb{1}(X^T\theta \leq 0, X^T\theta_0 \leq 0) - \mathbb{1}(X^T\theta_0 \leq 0)\right] \\ &+ (\beta_0 - \gamma_{mean})\mathbb{1}(X^T\theta \leq 0, X^T\theta_0 > 0)\big) \\ &= (\beta_0 - \gamma_{mean})P(X^T\theta \leq 0 < X^T\theta_0) - (\alpha_0 - \gamma_{mean})P(X^T\theta_0 \leq 0 < X^T\theta) \\ &\geq \max\left\{(\beta_0 - \gamma_{mean})P(X^T\theta \leq 0 < X^T\theta_0), (\gamma_{mean} - \alpha_0)P(X^T\theta_0 \leq 0 < X^T\theta)\right\} \\ &\geq a^- \frac{(\beta_0 - \alpha_0)}{2}\|\theta - \theta_0\|_2 \end{aligned}$$

which concludes the lemma. □

**Note:** In a similar way we can establish the upper bound on $R(\theta) - R(\theta^0)$ i.e.

$$R(\theta) - R(\theta^0) \leq a^+(\beta_0 - \alpha_0)\|\theta - \theta_0\|_2$$

using Assumption A1.

**Lemma 23.** $\|\hat{\theta} - \theta_0\|_2 \xrightarrow{P} 0$ *as* $n \to \infty$.

*Proof.*

$$
\begin{aligned}
P(\|\hat{\theta} - \theta_0\|_2 > \epsilon) &= P\left( \sup_{\|\theta-\theta_0\|_2>\epsilon} (\mathbb{M}_{n,d}(\theta_0) - \mathbb{M}_{n,d}(\theta)) \geq 0 \right) \\
&= P\Bigg( \sup_{\|\theta-\theta_0\|_2>\epsilon} (\mathbb{M}_{n,d}(\theta_0) - \mathbb{M}_{n,d}(\theta) - R_d(\beta_0) + R_d(\theta)) \\
&\quad + (R_d(\theta_0) - R_d(\theta)) \geq 0 \Bigg) \\
&= P\Bigg( \sup_{\|\theta-\theta_0\|_2>\epsilon} (\mathbb{M}_{n,d}(\theta_0) - \mathbb{M}_{n,d}(\theta) - R_d(\theta_0) + R_d(\theta)) \\
&\quad \geq \inf_{\|\theta-\theta_0\|_2>\epsilon} (R_d(\theta) - R_d(\theta_0)) \Bigg) \\
&\leq P\left( 2 \sup_{\|\theta-\theta_0\|_2>\epsilon} |\mathbb{M}_{n,d}(\theta) - R_d(\theta)|) \geq \inf_{\|\theta-\theta_0\|_2>\epsilon} (R_d(\theta) - R_d(\theta_0)) \right) \\
&\leq P\left( \sup_{\theta \in S^{d-1}} |\mathbb{M}_{n,d}(\theta) - R_d(\theta)|) \geq (a^-/2)(\beta_0 - \alpha_0)\epsilon \right) \\
&\xrightarrow{P} 0 \text{ as } n \to \infty
\end{aligned}
$$

which completes the proof. □

### B.0.1.2 Proof of Theorem III.5

In the previous section we have proved consistency of our estimator. We have also seen that the population criterion have the following curvature near the truth:

$$R(\theta_0) - R(\theta) \leq -a^-(\theta_0 - \alpha_0)\|\theta - \theta_0\|_2$$

To apply Theorem B.2 and obtain the rate of convergence, we consider the pseudo-distance function $d^\star(\theta, \theta_0) = \sqrt{\|\theta - \theta_0\|_2}$. This leaves the task of controlling for the

modulus of continuity, i.e. we need to find $\phi(\delta) \equiv \phi_{n,p}(\delta)$ for $\delta > 0$ small such that

$$\mathbb{E}\left(\sup_{d(\theta,\theta_0)\leq\delta} |\mathbb{M}_{d,n}(\theta) - \mathbb{M}_{d,n}(\theta_0) - R_d(\theta) + R_d(\theta_0)|\right) \leq \frac{\phi(\delta)}{\sqrt{n}}$$

and $\phi(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$. We can start with the following:

$$\begin{aligned}
&\mathbb{E}\left(\sup_{\|\theta-\theta_0\|_2\leq\delta^2} \sqrt{n}|(\mathbb{M}_{d,n} - R_d)(\theta-\theta_0)|\right) \\
&\leq \mathbb{E}\left(\sup_{\|\theta-\theta_0\|_2\leq\delta^2} \sqrt{n}|(L_{d,n} - R_d)(\theta-\theta_0)|\right) + \mathbb{E}\left(\sup_{\|\theta-\theta_0\|_2\leq\delta^2} \sqrt{n}|B_{d,n}(\theta-\theta_0)|\right) \\
&= S_1 + S_2 \qquad \text{(B.48)}
\end{aligned}$$

The first term can be controlled through a VC dimsntion argument. We have to show that, while the second term will not affect the modulus of continuity. Towards analyzing the second term:

$$\begin{aligned}
&\mathbb{E}\left(\sup_{\|\theta-\theta_0\|_2\leq\delta^2} \sqrt{n}|B_{d,n}(\theta-\theta_0)|\right) \\
&= \mathbb{E}\left(\sqrt{n}|\bar{Y} - \gamma_{mean}| \sup_{\|\theta-\theta_0\|_2\leq\delta^2} \left|\frac{1}{n}\sum_{i=1}^n \left\{\mathbb{1}(X^T\theta \leq 0) - \mathbb{1}(X^T\theta_0 \leq 0)\right\}\right|\right) \\
&\leq \sqrt{n}(\mathbb{E}(\bar{Y} - \gamma_{mean})^2)^{\frac{1}{2}} \left(\mathbb{E}\left(\sup_{\|\theta-\theta_0\|_2\leq\delta^2} \left|\frac{1}{n}\sum_{i=1}^n \left\{\mathbb{1}(X^T\theta \leq 0) - \mathbb{1}(X^T\theta_0 \leq 0)\right\}\right|\right)^2\right)^{\frac{1}{2}} \\
&= \sqrt{\gamma_{mean}(1-\gamma_{mean})} \left(\mathbb{E}\left(\sup_{\|\theta-\theta_0\|_2\leq\delta^2} \left|\frac{1}{n}\sum_{i=1}^n \left\{\mathbb{1}(X^T\theta \leq 0) - \mathbb{1}(X^T\theta_0 \leq 0)\right\}\right|^2\right)\right)^{\frac{1}{2}} \\
&\leq \sqrt{\gamma_{mean}(1-\gamma_{mean})} \left(\mathbb{E}\left(\sup_{\|\theta-\theta_0\|_2\leq\delta^2} \left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}\left(\text{sgn}(X^T\theta) \neq \text{sgn}(X^T\theta_0)\right)\right)^2\right)\right)^{\frac{1}{2}} \\
&\leq \sqrt{\gamma_{mean}(1-\gamma_{mean})} \left(\mathbb{E}\left(\sup_{\|\theta-\theta_0\|_2\leq\delta^2} \frac{1}{n}\sum_{i=1}^n \mathbb{1}\left(\text{sgn}(X^T\theta) \neq \text{sgn}(X^T\theta_0)\right)\right)\right)^{\frac{1}{2}} \\
&\leq \sqrt{\gamma_{mean}(1-\gamma_{mean})} \left[\mathbb{E}^{\frac{1}{2}}\left(\sup_{\|\theta-\theta_0\|_2\leq\delta^2} \frac{1}{n}\sum_{i=1}^n \left[\mathbb{1}\left(\text{sgn}(X^T\theta) \neq \text{sgn}(X^T\theta_0)\right)\right.\right.\right.
\end{aligned}$$

$$
\begin{aligned}
&- P\left(\operatorname{sgn}(X^T\theta) \neq \operatorname{sgn}(X^T\theta_0)\right)\Bigg]\Bigg) + \sqrt{\delta}\Bigg] \\
&\leq \sqrt{\gamma_{mean}(1-\gamma_{mean})}\ \left[\mathbb{E}^{\frac{1}{2}}\left(\sup_{\|\theta-\theta_0\|_2\leq\delta^2}(\mathbb{P}_n - P)(g_\theta)\right) + \delta\right] \\
&=: \sqrt{\gamma_{mean}(1-\gamma_{mean})}\ [T + \delta] \qquad \text{(B.49)}
\end{aligned}
$$

where $g_\theta(x) = \mathbb{1}(\operatorname{sgn}(X^T\theta) \neq \operatorname{sgn}(X^T\theta_0))$.

To bound the terms $S_1$ and $T$ we again refer back to Theorem B.3. With both terms we will be dealing with classes of functions which has VC dimensions bounded by some constant times $d$. From established results (see *der Vaart and Wellner* (1996)), we know that for a VC class of functions $\mathcal{F}$ with VC dimension $V$, a measurable envelope function $F$ and $r \geq 1$ one has for any probability measure $Q$ with $\|F\|_{Q,r} > 0$,

$$
N(\epsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \leq KV(4e)^V \left(\frac{2}{\epsilon}\right)^{rV} \qquad \text{(B.50)}
$$

for some universal constant $K$ and $0 < \epsilon < 1$. In turn this means

$$
N(\epsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq \left(\frac{A}{\epsilon}\right)^{2V}
$$

where $A \geq \left(e^{\frac{\log k + \log V}{2V}} 2\sqrt{e}\right) \vee e^2$, which can be bounded above by a $n$-independent constant greater than $e^2$, even should $V$ grows with $n$. It remains to apply Theorem B.3 to the function classes associated with $S_1$ and $T$.

**Upper bounding $S_1$ from equation (B.48):** Here our collection of functions is:

$$
\mathcal{F}_\delta = \{f_\theta : f_\theta(x,y) = (y-\gamma)\left(\mathbb{1}_{x^T\theta\leq 0} - \mathbb{1}_{x^T\theta_0\leq 0}\right), \|\theta - \theta_0\|_2 \leq \delta\},
$$

which has a VC dimension bounded by a universal constant times $d$. For any such $f_\theta$:

$$\begin{aligned}\mathrm{Var}(f_\theta(X,Y)) &\leq \mathbb{E}(f_\theta^2(X,Y)) \\ &\leq \mathbb{E}\left(\mathbb{1}_{X^T\theta\leq 0} - \mathbb{1}_{X^T\theta_0\leq 0}\right)^2 \\ &= P(\mathrm{sgn}(X^T\theta) \neq \mathrm{sgn}(X^T\theta_0)) \\ &\leq a^+\|\theta-\theta_0\|_2\end{aligned}$$

Hence we have

$$\sup_{f_\theta\in\mathcal{F}_\delta} \mathrm{Var}(f_\theta(X,Y)) \leq a^+\delta^2.$$

Apply Theorem B.3 to obtain:

$$\begin{aligned}&\mathbb{E}\left(\sup_{\|\theta-\theta_0\|_2\leq\delta^2} \sqrt{n}|(L_{d,n}-R_d)(\theta-\theta_0)|\right) \\ &= \frac{1}{\sqrt{n}}\mathbb{E}\left(\sup_{\|\theta-\theta_0\|_2\leq\delta^2}\left|\sum_{i=1}^n\left[(Y_i-\gamma)\mathbb{1}_{X_i^T\theta\leq 0} - P\left((Y-\gamma)\mathbb{1}_{X^T\theta\leq 0}\right)\right]\right|\right) \\ &\lesssim \frac{1}{\sqrt{n}}\left(\delta\sqrt{nd\log\frac{1}{\delta}} \vee d\log\frac{1}{\delta}\right) \\ &\lesssim \delta\sqrt{d\log\frac{1}{\delta}} \vee \frac{d}{\sqrt{n}}\log\frac{1}{\delta}\end{aligned}$$

**Upper bounding $T$ from equation 12:** To control $T$, we focus on the set of functions:

$$\mathcal{G}_\delta = \{g_\theta : g_\theta(x) = \mathbb{1}(\mathrm{sgn}(x^T\theta) \neq \mathrm{sgn}(x^T\theta_0)), \|\theta-\theta_0\|_2 \leq \delta\}$$

Here the variance factor is:

$$\begin{aligned}\mathrm{Var}(g_\theta(X)) &\leq \mathbb{E}(\mathbb{1}(\mathrm{sgn}(X^T\theta) \neq \mathrm{sgn}(X^T\theta_0))) \\ &= P(\mathrm{sgn}(X^T\theta) \neq \mathrm{sgn}(X^T\theta_0)) \\ &\leq a^+\|\theta - \theta_0\|_2\end{aligned}$$

This is also a class of functions with VC dimension bounded by a universal constant times $d$. We also have the bound:

$$\sup_{g_\theta \in \mathcal{G}_\delta} \mathrm{Var}(g_\theta(X)) \leq a^+\delta^2$$

for all $g_\theta$ within the class $\mathcal{G}_\delta$. Using this we can derive the following:

$$\begin{aligned}T^2 &= \mathbb{E}\left(\sup_{\|\theta-\theta_0\|_2 \leq \delta^2} |\mathbb{P}_n - P|(g_\theta)\right) \\ &= \frac{1}{n}\mathbb{E}\left(\sup_{\|\theta-\theta_0\|_2 \leq \delta^2} \left|\sum_{i=1}^n (g_\theta(X_i) - Pg_\theta)\right|\right) \\ &\lesssim \frac{1}{n}\left(\delta\sqrt{nd\log\frac{1}{\delta}} \vee d\log\frac{1}{\delta}\right) \\ &\lesssim \frac{\delta}{\sqrt{n}}\sqrt{d\log\frac{1}{\delta}} \vee \frac{d}{n}\log\frac{1}{\delta}\end{aligned}$$

which implies

$$T \lesssim \sqrt{\frac{\delta}{\sqrt{n}}\sqrt{d\log\frac{1}{\delta}} \vee \frac{d}{n}\log\frac{1}{\delta}}$$

Combining these upper bounds and using the inequalities:

1. $a \vee b \leq a + b$ when $a, b, \geq 0$
2. $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$

we have:

$$\phi_{n.d}(\delta) = K\left[\delta\sqrt{d\log\frac{1}{\delta}} + \frac{d}{\sqrt{n}}\log\frac{1}{\delta} + \sqrt{\frac{\delta}{\sqrt{n}}\sqrt{d\log\frac{1}{\delta}}} + \sqrt{\frac{d}{n}\log\frac{1}{\delta}} + \delta\right]$$

for some universal constant $K$. This is the sum of 5 terms, so to satisfy the inequality

$$r_n^2\phi_n(1/r_n) \leq \sqrt{n},$$

we separately consider the cases where the first, second, ..., firth term is the maximum term.

1. Taking the first term to be maximal would mean $\phi(\delta) \lesssim \delta\sqrt{d\log\frac{1}{\delta}}$, leading to the rate $r_n = \left(\frac{n}{d}\right)^{\frac{1}{2}}\left(\log\frac{n}{d}\right)^{-\frac{1}{2}}$.

2. Taking the second term as maximal means $\phi(\delta) \lesssim \frac{d}{\sqrt{n}}\log\frac{1}{\delta}$, we get $r_n = \left(\frac{n}{d}\right)^{\frac{1}{2}}\left(\log\frac{n}{d}\right)^{-\frac{1}{2}}$.

3. Taking $\phi(\delta) \lesssim \sqrt{\frac{\delta}{\sqrt{n}}\sqrt{d\log\frac{1}{\delta}}}$ we have $r_n = \left(\frac{n^3}{d}\right)^{\frac{1}{6}}\left(\log\frac{n^3}{d}\right)^{-\frac{1}{6}}$.

4. Taking $\phi(\delta) \lesssim \sqrt{\frac{d}{n}\log\frac{1}{\delta}}$ we have $r_n = \left(\frac{n^2}{d}\right)^{\frac{1}{4}}\left(\log\frac{n^2}{d}\right)^{-\frac{1}{4}}$.

5. Taking $\phi(\delta) \lesssim \delta$, we have $r_n = n$.

Hence the slowest rate (corresponding to the largest $r_n$) is $\left(\frac{n}{d}\right)^{\frac{1}{2}}\left(\log\frac{n}{d}\right)^{-\frac{1}{2}}$ if $d^2\left(\log n/d\right)^3 \geq \log n^3/d$ and $d(\log(n/d))^2 \geq \log(n^2/d)$ for all large $n$. To show that these are indeed the case, we note that the first condition can be re-written as

$$\begin{aligned} d^2\left(\log n/d\right)^3 \geq \log n^3/d &\iff d^2\left(\log n/d\right)^3 \geq \log n^3/d^3 + 2\log d \\ &\iff d \geq \frac{3\log n/d}{(\log n/d)^3} + 2\frac{\log d}{d(\log n/d)^3} \\ &\iff d \geq \frac{3}{(\log n/d)^2} + 2\frac{\log d}{d(\log n/d)^3} \end{aligned}$$

which is trivially true as both the terms on the RHS is going to 0 as $n, d \to \infty$ with $n/d \to 0$. Similarly,

$$d\left(\log\frac{n}{d}\right)^2 \geq \log\frac{n^2}{d} \iff d\left(\log\frac{n}{d}\right)^2 \geq 2\log\frac{n}{d} + \log d$$
$$\iff d \geq \frac{2}{\log\frac{n}{d}} + \frac{\log d}{\left(\log\frac{n}{d}\right)^2}$$

which is also trivially true as the first on the RHS of the bottom line goes to 0, while the second term is strictly less than $\log d$, which is less than $d$. Hence we have proved:

$$d^\star(\theta, \theta_0) = O_p\left(\left(\frac{d}{n}\right)^{\frac{1}{2}}\left(\log\frac{n}{d}\right)^{\frac{1}{2}}\right)$$

Using the fact that $d^\star(\theta, \theta_0)^2 = \|\theta - \theta_0\|_2$ we have

$$\|\theta - \theta_0\|_2 = O_p\left(\left(\frac{d}{n}\right)\left(\log\frac{d}{n}\right)\right).$$

### B.0.2 Proofs for Section 3.5

In the next several sections we will repeatedly refer to the following random and deterministic functions

1. $\mathbb{M}^*_{n,d}(\theta) = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})\left(\mathbb{1}(X_i^T\theta \leq 0) - \mathbb{1}(X_i^T\theta_0 \leq 0)\right)$
2. $L^*_{n,d}(\theta) = \frac{1}{n}\sum_{i=1}^n (Y_i - \gamma_{mean})\left(\mathbb{1}(X_i^T\theta \leq 0) - \mathbb{1}(X_i^T\theta_0 \leq 0)\right)$
3. $B^*_{n,d}(\theta) = (\bar{Y} - \gamma_{mean})\frac{1}{n}\sum_{i=1}^n \left(\mathbb{1}(X_i^T\theta \leq 0) - \mathbb{1}(X_i^T\theta_0 \leq 0)\right)$
4. $R^*_d = P(Y - \gamma_{mean})\left(\mathbb{1}(X^T\theta \leq 0) - \mathbb{1}(X^T\theta_0 \leq 0)\right)$

The random function $\mathbb{M}^*_{n,d}$ relates to the estimator $\hat{\theta}_{\mathcal{H}}$, the latter being the minimizer of the former:

$$\hat{\theta}_{\mathcal{H}} := \arg\min_{\theta\in\mathcal{H}} \mathbb{M}_{n,d}(\theta) \tag{B.51}$$

for any $\mathcal{H} \subseteq S^{d-1}$ of the form $\mathcal{H} = \{\theta \in S^{d-1} : \|\theta\|_0 \leq d_0\}$. Similarly, the parameter $\theta_0$ can be shown to be the minimizer of $R_d^*$. The proofs we will show in this subsection will make heavy use of these notations to argue that $\hat{\theta}_{\mathcal{H}}$ converges to $\theta_0$ in terms of Euclidean distance and support.

#### B.0.2.1 Preliminary Results for this Section

The proofs of results from the later parts of Section 3.5 will refer to two other main results, the first being a bound for the VC dimension of the estimating function classes, and the second being probability bounds on the excess risk and empirical loss.

We first focus on upper bounding the VC dimension of the parameter space $\mathcal{H}$.

**Lemma 24.** *The class of functions* $\mathcal{F}_{d_0}^* := \{f(x,y) = (y - \gamma_{mean}) \cdot \mathbb{1}(x^T\theta \leq 0) : \theta \in \mathcal{H}\}$ *has a VC dimension of at most* $Kd_0 \log\left(\frac{ed}{d_0}\right)$ *for some universal constant* $K > 1$.

*Proof.* For each $m \subseteq \{1, \ldots, d\}$ such that $|m| = d_0$, we look at the class of functions

$$\mathcal{F}_{d_0}^*(m) := \{f(x,y) = (y - \gamma_{mean}) \cdot \mathbb{1}(x^T\theta \leq 0) : \theta \in \mathcal{H},\ \theta_j = 0 \text{ if } j \notin m\}$$

The class of functions $\mathcal{F}_{d_0}^*$ is the union of all classes $\mathcal{F}_{d_0}^*(m)$. Denote the VC dimension of $\mathcal{F}_{d_0}^*(m)$ to be $v(m)$; it is known that $v(m) \leq Ld_0$ for some universal constant $L$. Denote the subgraph classes of $\mathcal{F}_{d_0}^*$ and $\mathcal{F}_{d_0}^*(m)$ as

$$\begin{aligned} SG &:= \left\{\{(x,y,z) \in S^{d-1} \times \{0,1\} \times \mathbb{R} : z < f(x,y)\} : f \in \mathcal{F}_{d_0}^*\right\} \\ SG_m &:= \left\{\{(x,y,z) \in S^{d-1} \times \{0,1\} \times \mathbb{R} : z < f(x,y)\} : f \in \mathcal{F}_{d_0}^*(m)\right\}. \end{aligned} \tag{B.52}$$

By definition, $SG$ has the same VC dimension as $\mathcal{F}_{d_0}^*$ (denote this VC dimension as $V$), and each $SG_m$ has VC dimension $v(m)$. Also, $SG$ equals the union of all $SG_m$'s.

Let $W := \{w_1, \ldots, w_{V-1}\}$ be any set in $S^{d-1} \times \{0,1\} \times \mathbb{R}$ which is shattered by $SG$. On the other hand, for every $m \subseteq \{1, \ldots, d\}$, the number of subsets of $W$ picked out by $SG_m$ is at most

$$\begin{aligned} \sum_{k=0}^{v(m)-1} \binom{V-1}{k} &\leq \left(\frac{e(V-1)}{v(m)-1}\right)^{v(m)-1} \\ &\leq \left(\frac{e(V-1)}{Ld_0}\right)^{d_0+1}, \end{aligned} \tag{B.53}$$

where the first expression is due to Sauer's Lemma, and the last inequality is due to $f(x) = (A/x)^x$ being an increasing function whenever $0 < x < (A/e)$. From here, note that the number of subsets of $W$ picked out by $SG$ (which is $2^{V-1}$) must not exceed the summation of the number of subsets of $W$ picked out by each $SG_m$, and therefore

$$\begin{aligned} 2^{V-1} &\leq \sum_m \left(\frac{e(V-1)}{d_0+1}\right)^{Ld_0} \\ &\leq \binom{d}{d_0}\left(\frac{e(V-1)}{d_0+1}\right)^{Ld_0} \\ &\leq \left(\frac{ed}{d_0}\right)^{d_0}\left(\frac{e(V-1)}{d_0+1}\right)^{Ld_0} \\ &\leq \left(\frac{ed}{d_0}\right)^{Ld_0}\left(\frac{ed}{d_0}\right)^{Ld_0}. \end{aligned} \tag{B.54}$$

This leads to the conclusion that

$$V \leq 2Ld_0 \log_2\left(\frac{ed}{d_0}\right) + 1 \leq (2L+1)d_0 \log_2\left(\frac{ed}{d_0}\right) \tag{B.55}$$

□

Next we will state a result which will give us probability bounds on the excess risk and empirical loss. In particular this result will be useful for deriving Theorem III.6. This proof of this preliminary result is very similar to the proof of Theorem 9.1

in *Giraud* (2014), so details regarding its derivation will be placed later on in Section B.0.2.4.

**Theorem B.6.** *Let $\mathcal{I}$ be a subset of $S^{d-1}$ and define*

$$\hat{\theta}_{\mathcal{I}} := \arg\min_{\theta \in \mathcal{I}} \mathbb{M}^*_{n,d}(\theta). \tag{B.56}$$

*For any $s > 0$, with probability greater than $1 - e^{-s}$:*

$$\begin{aligned} \left| R^*_d(\hat{\theta}_{\mathcal{I}}) - \min_{\theta \in \mathcal{I}} R^*_d(\theta) \right| &\leq 2K\sqrt{\frac{V_n}{n}} + 2\sqrt{\frac{1}{n}} + 2\sqrt{\frac{2s}{n}} + \frac{4}{n} \\ \left| R^*_d(\hat{\theta}_{\mathcal{I}}) - \mathbb{M}^*_{n,d}(\hat{\theta}_{\mathcal{I}}) \right| &\leq K\sqrt{\frac{V_n}{n}} + \sqrt{\frac{1}{n}} + \sqrt{\frac{2s}{n}} + \frac{2}{n} \end{aligned} \tag{B.57}$$

*where $V_n$ is the VC dimension of $\mathcal{I}$ and $K$ is a universal constant.*

### B.0.2.2 Proof of Theorem III.6

We re-iterate the result before moving on to its proof. Suppose $m_1$ and $m_2$ are subsets of $\{1. \ldots, d\}$ with $d_0$ elements each, such that $supp(\theta_0) \subseteq m_2$ and $supp(\theta_0) \nsubseteq m_1$, then

$$\hat{\theta}_{m_j} := \underset{\substack{\theta \in S^{d-1} \\ \theta_k = 0 \text{ for } k \notin m_j}}{\arg\min} \; \mathbb{M}_{n,d}(\theta) \tag{B.58}$$

for $j = 1, 2$. Then there exists a constant $C > 0$ such that if the following inequality is satisfied:

$$\theta_{0,min} := \min_{j:|\theta_{0,j}| \neq 0} |\theta_{0,j}| \geq C\sqrt{\frac{d_0 \log\left(\frac{d}{d_0}\right)}{n}}, \tag{B.59}$$

then with probability at least $1 - 2\exp\left(-Ld_0 \log\left(\frac{d}{d_0}\right)\right)$,

$$\mathbb{M}_{n,d}(\hat{\theta}_{m_2}) < \mathbb{M}_{n,d}(\hat{\theta}_{m_2}) \tag{B.60}$$

for some constant $L > 2$ not dependent on $n$.

*Proof.* Define

$$\mathbb{M}^*_{n,d}(\theta) := \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})\left(\mathbb{1}(X_i^T\theta \leq 0) - \mathbb{1}(X_i^T\theta_0 \leq 0)\right) \tag{B.61}$$

and note that $\hat{\theta}_{m_j}$ are also equal to

$$\hat{\theta}_{m_j} := \underset{\substack{\theta\in S^{d-1}\\ \theta_k=0 \text{ for } k\notin m_j}}{\arg\min} \ \mathbb{M}^*_{n,d}(\theta). \tag{B.62}$$

Also define $R^*_d := P(Y - \gamma_{mean})\left(\mathbb{1}(X^T\theta \leq 0) - \mathbb{1}(X^T\theta_0 \leq 0)\right)$, and similarly let

$$\tilde{\theta}_{m_j} := \underset{\substack{\theta\in S^{d-1}\\ \theta_k=0 \text{ for } k\notin m_j}}{\arg\min} \ R^*_d(\theta). \tag{B.63}$$

Because the support of $\theta_0$ resides within $m_2$, we must have $\tilde{\theta}_{m_2} = \theta_0$. Therefore the difference between the values of $R^*_d(\tilde{\theta}_{m_1})$ and $R^*_d(\tilde{\theta}_{m_2})$ can be bounded by

$$\begin{aligned}
R^*_d(\hat{\theta}_{m_1}) - R^*_d(\hat{\theta}_{m_2}) &= P(Y - \gamma_{mean})\left(\mathbb{1}(X^T\tilde{\theta}_{m_1} \leq 0) - \mathbb{1}(X^T\theta_0 \leq 0)\right) \\
&= -(\alpha_0 - \gamma_{mean})P\Big(X^T\theta_0 \leq 0 < X^T\tilde{\theta}_{m_1}\Big) \\
&\quad +(\beta_0 - \gamma_{mean})P\Big(X^T\theta_0 > 0 \geq X^T\tilde{\theta}_{m_1}\Big) \\
&\geq a^- \max\Big\{(\gamma_{mean} - \alpha_0), (\beta_0 - \gamma_{mean})\Big\}\|\tilde{\theta}_{m_1} - \theta_0\|_2 \\
&\geq a^- \max\Big\{(\gamma_{mean} - \alpha_0), (\beta_0 - \gamma_{mean})\Big\}\theta_{min} \\
&\quad \text{because } m_1 \text{ does not contain at least one entry from } supp(\theta_{0,min})
\end{aligned} \tag{B.64}$$

Using the probability bounds from Theorem B.6 and the VC dimension bound

from Theorem 24, we have

$$\begin{aligned}\left|\mathbb{M}^*_{n,d}(\hat{\theta}_{m_1}) - R^*_d(\tilde{\theta}_{m_1})\right| &\leq \left|\mathbb{M}^*_{n,d}(\hat{\theta}_{m_1}) - R^*_d(\hat{\theta}_{m_1})\right| + \left|R^*_d(\hat{\theta}_{m_1}) - R^*_d(\tilde{\theta}_{m_1})\right| \\ &\leq 3K\sqrt{\frac{Ld_0\log\left(\frac{ed}{d_0}\right)}{n}} + 9\sqrt{\frac{1}{n}} + 3\sqrt{\frac{2s}{n}} \qquad \text{(B.65)}\end{aligned}$$

with probability greater than $1 - e^{-s}$, for any $s > 0$, and for some universal constant $K$ and $L > 2$. By picking $s = \frac{L}{2}d_0\log\left(\frac{ed}{d_0}\right)$ we have

$$\left|\mathbb{M}^*_{n,d}(\hat{\theta}_{m_1}) - R^*_d(\tilde{\theta}_{m_1})\right| \leq (3K+12)\sqrt{\frac{Ld_0\log\left(\frac{ed}{d_0}\right)}{n}} \qquad \text{(B.66)}$$

with probability at least $1 - \exp\left(-\frac{L}{2}d_0\log\left(\frac{ed}{d_0}\right)\right)$. Similarly, deduce that with probability at least $1 - \exp\left(-\frac{L}{2}d_0\log\left(\frac{ed}{d_0}\right)\right)$,

$$\left|\mathbb{M}^*_{n,d}(\hat{\theta}_{m_2}) - R^*_d(\tilde{\theta}_{m_2})\right| \leq (3K+12)\sqrt{\frac{Ld_0\log\left(\frac{ed}{d_0}\right)}{n}}. \qquad \text{(B.67)}$$

Now suppose

$$\theta_{min} > \left(a^- \max\left\{(\gamma_{mean} - \alpha_0), (\beta_0 - \gamma_{mean})\right\}\right)^{-1}(3K+12)\sqrt{\frac{Ld_0\log\left(\frac{ed}{d_0}\right)}{n}}, \qquad \text{(B.68)}$$

then with probability at least $1 - 2\exp\left(-\frac{L}{2}d_0\log\left(\frac{ed}{d_0}\right)\right)$, we have

$$\begin{aligned}&\mathbb{M}^*_{n,d}(\hat{\theta}_{m_1}) \\ \geq\ & R^*_d(\tilde{\theta}_{m_1}) - (3K+12)\sqrt{\frac{Ld_0\log\left(\frac{ed}{d_0}\right)}{n}} \\ \geq\ & R^*_d(\tilde{\theta}_{m_2}) + a^- \min\left\{(\gamma_{mean} - \alpha_0), (\beta_0 - \gamma_{mean})\right\}\theta_{min}\end{aligned}$$

$$
\begin{aligned}
& -(3K+12)\sqrt{\frac{Ld_0\log\left(\frac{ed}{d_0}\right)}{n}} \\
> \; & R_d^*(\tilde{\theta}_{m_2}) + a^- \min\Big\{(\gamma_{mean}-\alpha_0),(\beta_0-\gamma_{mean})\Big\}\theta_{min} \\
& -(3K+12)\sqrt{\frac{Ld_0\log\left(\frac{ed}{d_0}\right)}{n}} \\
\geq \; & \mathbb{M}_{n,d}^*(\hat{\theta}_{m_2}).
\end{aligned} \tag{B.69}
$$

Adding $\frac{1}{n}\sum_i (Y_i-\bar{Y})\mathbb{1}(X_i^T\theta_0\leq 0)$ to both sides would obtain $\mathbb{M}_{n,d}(\hat{\theta}_{m_1}) > \mathbb{M}_{n,d}(\hat{\theta}_{m_2})$.

□

#### B.0.2.3 Proof of Theorem III.8

We will first prove that $\|\hat{\theta}_{\mathcal{H}}-\theta_0\|_2$ converges to 0 in probability.

*Proof.* First note that

$$
\sup_{\theta\in\mathcal{H}} |\mathbb{M}_{n,d}(\theta) - R_d(\theta)| \to 0 \tag{B.70}
$$

in probability. This can be established by separating into two terms which goes to 0 in probability

$$
\begin{aligned}
\sup_{\theta\in\mathcal{H}} |\mathbb{M}_{n,d}(\theta) - R_d^*(\theta)| \; \leq \; & \sup_{\theta\in\mathcal{H}} \left|(\mathbb{P}_n - P)(Y-\gamma_{mean})\mathbb{1}(X^T\theta\leq 0)\right| \\
& + \sup_{\theta\in\mathcal{H}} \left|(\bar{Y}-\gamma_{mean})\frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i^T\theta\leq 0)\right| \\
\leq \; & \sup_{\theta\in\mathcal{H}} \left|(\mathbb{P}_n - P)(Y-\gamma_{mean})\mathbb{1}(X^T\theta\leq 0)\right| + \left|\bar{Y}-\gamma_{mean}\right|.
\end{aligned} \tag{B.71}
$$

The second term $\left|\bar{Y}-\gamma_{mean}\right|$ converges to 0 in probability due to the fact that $Y$'s are marginally binomial with mean $\gamma_{mean}$, and hence $(\bar{Y}-\gamma_{mean})$ has a variance no greater than $\frac{1}{4n}$. As for the first term, the class $\mathcal{F}_{d_0}^*$ has an envelope function of $F_n(x,y)=1$

for all $n$. Using the method of Theorem 2.4.3 from *der Vaart and Wellner* (1996), we have

$$\begin{aligned} & \sup_{\theta\in\mathcal{H}} \left|(\mathbb{P}_n - P)(Y-\gamma_{mean})\mathbb{1}(X^T\theta \le 0)\right| \\ \lesssim & \frac{1}{\sqrt{n}}\sqrt{1+\log(N(\varepsilon, \mathcal{F}^*_{d_0}, L_1(\mathbb{P}_n)))} \\ \lesssim & \sqrt{\frac{d_0 \log\left(\frac{ed}{d_0}\right)}{n}} \\ \to & 0, \end{aligned} \tag{B.72}$$

thus showing (B.70).

From Lemma 22, there exists a constant $K > 0$, not dependent on $n$, such that for any $\theta \in \mathcal{H}$,

$$R_d(\theta) - R_d(\theta_0) \ge K\|\theta - \theta_0\|_2. \tag{B.73}$$

Because $\hat{\theta}_{\mathcal{H}}$ is the minimizer of $\mathbb{M}_{n,d}(\theta)$, this means that for any $\varepsilon > 0$, we have

$$\begin{aligned} & \mathbb{P}\left(\|\hat{\theta}_{\mathcal{H}} - \theta_0\|_2 > \varepsilon\right) \\ \le & \mathbb{P}\left(R_d(\hat{\theta}_{\mathcal{H}}) - R_d(\theta_0) > K\varepsilon\right) \\ \le & \mathbb{P}\left(R_d(\hat{\theta}_{\mathcal{H}}) - \mathbb{M}_{n,d}(\hat{\theta}_{\mathcal{H}}) - R_d(\theta_0) + \mathbb{M}_{n,d}(\theta_0) > K\varepsilon\right) \\ \le & \mathbb{P}\left[\sup_{\theta\in\mathcal{H}} \left|(\mathbb{P}_n - P)(Y-\gamma_{mean})\mathbb{1}(X^T\theta \le 0)\right| > \frac{K\varepsilon}{2}\right] \\ \to & 0 \end{aligned} \tag{B.74}$$

□

Next we shall prove the rate of convergence of $\hat{\theta}_{\mathcal{H}}$ to $\theta_0$.

*Proof.* Again we utilize to Theorem B.2. For this problem our associated distance

function would be the square root of the Euclidean norm, as Lemma 22 shows that

$$R_d^*(\theta_0) - R_d^*(\theta) \lesssim -\|\theta - \theta_0\|_2 \tag{B.75}$$

for all $\theta$ within a neighborhood of $\theta_0$. Our strategy would be to calculate

$$\mathbb{E}\left[\sup_{\substack{\|\theta-\theta_0\|_2\le\delta^2\\ \theta\in\mathcal{H}}} \sqrt{n}\left|(\mathbb{M}_{n,d}(\theta) - R_d(\theta)) - (\mathbb{M}_{n,d}(\theta_0) - R_d(\theta_0))\right|\right]. \tag{B.76}$$

We again bound this above by the two terms

$$\sqrt{n}\mathbb{E}\left[\sup_{f\in\mathcal{G}_\delta}\left|(\mathbb{P}_n - P)g\right|\right] + \mathbb{E}\left[\sup_{\substack{\|\theta-\theta_0\|_2\le\delta^2\\ \theta\in\mathcal{H}}} \sqrt{n}\left|(\bar{Y} - \gamma_{mean})(\mathbb{1}(X^T\theta \le 0) - \mathbb{1}(X^T\theta_0 \le 0))\right|\right]. \tag{B.77}$$

where $\mathcal{G}_\delta := \{f(x,y) = (y-\gamma_{mean})(\mathbb{1}(x^T\theta \le 0) - \mathbb{1}(x^T\theta_0 \le 0)) : \theta \in \mathcal{H}, \|\theta - \theta_0 \le \delta^2\}$. The second term can be bounded as

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\substack{\|\theta-\theta_0\|_2\le\delta^2\\ \theta\in\mathcal{H}}} \sqrt{n}\left|(\bar{Y} - \gamma_{mean})(\mathbb{1}(X^T\theta \le 0) - \mathbb{1}(X^T\theta_0 \le 0))\right|\right] \\
\le\ & \sqrt{\mathbb{E}\left[\sup_{\|\theta-\theta_0\|_2\le\delta^2,\theta\in\mathcal{H}} (\bar{Y} - \gamma_{mean})^2\right]}\sqrt{\mathbb{E}\left[\sup_{\|\theta-\theta_0\|_2\le\delta^2,\theta\in\mathcal{H}} (\mathbb{1}(X^T\theta \le 0) - \mathbb{1}(X^T\theta_0 \le 0))^2\right]} \\
\le\ & \sqrt{\gamma_{mean}(1-\gamma_{mean})}\sqrt{\mathbb{E}\left[\sup_{\|\theta-\theta_0\|_2\le\delta^2,\theta\in\mathcal{H}} \mathbb{1}(\mathrm{sgn}(X^T\theta) \ne \mathrm{sgn}(X^T\theta_0))\right]} \\
\le\ & \sqrt{\gamma_{mean}(1-\gamma_{mean})}\sqrt{\mathbb{E}\left[\sup_{g\in\mathcal{J}_\delta} (\mathbb{P}_n - P)g\right] + \sup_{g\in\mathcal{J}_\delta} Pg} \\
& \text{where } \mathcal{J}_\delta := \{g(x,y) = \mathbb{1}(\mathrm{sgn}(x^T\theta) \ne \mathrm{sgn}(x^T\theta_0)) : \|\theta - \theta_0\|_2 \le \delta^2, \theta \in \mathcal{H}\} \\
\le\ & \sqrt{\gamma_{mean}(1-\gamma_{mean})}\sqrt{\mathbb{E}\left[\sup_{g\in\mathcal{J}_\delta} (\mathbb{P}_n - P)g\right] + a^+\delta^2}
\end{aligned} \tag{B.78}$$

To bound the value of $\mathbb{E}\left[\sup_{g\in\mathcal{J}_\delta}(\mathbb{P}_n - P)g\right]$ and $\mathbb{E}\left[\sup_{g\in\mathcal{G}_\delta}\left|(\mathbb{P}_n - P)g\right|\right]$, we refer to Theorem B.3. Both $\mathcal{G}_\delta$ and $\mathcal{J}_\delta$ have the envelope function $F(x,y) = 1$, and using

the logic of Lemma 24, both have VC dimensions bounded by $Cd_0 \log\left(\frac{ed}{d_0}\right)$ for some universal constant $C > 1$. By standard empirical process theory, there exists a universal constant $K$ such that

$$N(\epsilon U, \mathcal{F}, L_2(Q)) \leq K\left(Cd_0 \log\left(\frac{ed}{d_0}\right)\right)\left(\frac{16e}{\epsilon}\right)^{2Cd_0 \log\left(\frac{ed}{d_0}\right)} \tag{B.79}$$

for all $\epsilon \in (0,1)$, and therefore it is possible to fund some constant $A > e^2$ such that $(A/\epsilon)^{3Cd_0 \log\left(\frac{ed}{d_0}\right)}$ bounds the above expression for all $\epsilon \in (0,1)$. Furthermore, the value of $\sup_{f\in\mathcal{F}} P(f - Pf)^2 \leq \sup_{f\in\mathcal{F}} Pf^2$, for both $\mathcal{F} = \mathcal{G}_\delta$ or $\mathcal{F}_\delta$, can be bounded by

$$\sup_{\substack{\|\theta-\theta_0\|_2\leq\delta^2 \\ \theta\in\mathcal{H}}} P\mathbb{1}(\text{sgn}(X^T\theta) \neq \text{sgn}(X^T\theta_0)) = \sup_{\substack{\|\theta-\theta_0\|_2\leq\delta^2 \\ \theta\in\mathcal{H}}} \mathbb{P}(\text{sgn}(X^T\theta) \neq \text{sgn}(X^T\theta_0)) \leq a^+\delta^2 \tag{B.80}$$

Therefore, both $\mathbb{E}\left[\sup_{g\in\mathcal{J}_\delta}(\mathbb{P}_n - P)g\right]$ and $\mathbb{E}\left[\sup_{g\in\mathcal{G}_\delta}|(\mathbb{P}_n - P)g|\right]$ can be bounded by

$$\frac{C^*}{n}\left(\sqrt{n}\delta\sqrt{d_0 \log\left(\frac{ed}{d_0}\right)\log\left(\frac{A}{\delta}\right)} \vee d_0 \log\left(\frac{ed}{d_0}\right)\log\left(\frac{A}{\delta}\right)\right) \tag{B.81}$$

for some constant $C^*$ not dependent on $n$. Therefore, expression (B.78) can be bounded by a constant multiple (not dependent on $n$) of

$$\sqrt{\frac{1}{\sqrt{n}}\delta\sqrt{d_0 \log\left(\frac{ed}{d_0}\right)\log\left(\frac{A}{\delta}\right)} \vee \frac{d_0}{n}\log\left(\frac{ed}{d_0}\right)\log\left(\frac{A}{\delta}\right)} + \delta. \tag{B.82}$$

The expression in (B.76) can be bounded by a ($n$-independent) constant multiple of

$$\phi_{n,d}(\delta) = \delta\sqrt{d_0 \log\left(\frac{ed}{d_0}\right)\log\left(\frac{A}{\delta}\right)} \vee \frac{d_0}{\sqrt{n}}\log\left(\frac{ed}{d_0}\right)\log\left(\frac{A}{\delta}\right)$$

$$+\sqrt{\frac{1}{\sqrt{n}}\delta\sqrt{d_0\log\left(\frac{ed}{d_0}\right)\log\left(\frac{A}{\delta}\right)} \vee \frac{d_0}{n}\log\left(\frac{ed}{d_0}\right)\log\left(\frac{A}{\delta}\right)} + \delta. \tag{B.83}$$

Depending on which term is dominant, the inequality $r_n^2\phi_{n,d}(1/r_n) \leq \sqrt{n}$ will provide different results on $r_n$:

- if $\delta\sqrt{d_0\log\left(\frac{ed}{d_0}\right)\log\left(\frac{A}{\delta}\right)}$ is the largest term, then $r_n \sim \sqrt{\frac{n}{d_0\log\left(\frac{ed}{d_0}\right)}}\left(\log\left(\frac{n}{d_0\log\left(\frac{ed}{d_0}\right)}\right)\right)^{-1/2}$
- if $\frac{d_0}{\sqrt{n}}\log\left(\frac{ed}{d_0}\right)\log\left(\frac{A}{\delta}\right)$ is the largest term, then $r_n \sim \sqrt{\frac{n}{d_0\log\left(\frac{ed}{d_0}\right)}}\left(\log\left(\frac{n}{d_0\log\left(\frac{ed}{d_0}\right)}\right)\right)^{-1/2}$
- if $\sqrt{\frac{1}{\sqrt{n}}\delta\sqrt{d_0\log\left(\frac{ed}{d_0}\right)\log\left(\frac{A}{\delta}\right)}}$ is largest, then $r_n \sim \sqrt{\frac{n}{\sqrt[3]{d_0\log\left(\frac{ed}{d_0}\right)}}}\left(\log\left(\frac{n^3}{d_0\log\left(\frac{ed}{d_0}\right)}\right)\right)^{-1/6}$
- if $\sqrt{\frac{d_0}{n}\log\left(\frac{ed}{d_0}\right)\log\left(\frac{A}{\delta}\right)}$ is largest, then $r_n \sim \sqrt{\frac{n}{\sqrt{d_0\log\left(\frac{ed}{d_0}\right)}}}\left(\log\left(\frac{n^2}{d_0\log\left(\frac{ed}{d_0}\right)}\right)\right)^{-1/4}$
- if $\delta$ is largest, the $r_n \sim \sqrt{n}$

Because the smallest order of $r_n$ listed is $r_n \sim \sqrt{\frac{n}{d_0\log\left(\frac{ed}{d_0}\right)}}\left(\log\left(\frac{n}{d_0\log\left(\frac{ed}{d_0}\right)}\right)\right)^{-1/2}$. Along with the associated distance between the square root of the Euclidean norm, this leads to the conclusion that

$$\|\hat{\theta}_{\mathcal{H}} - \theta_0\|_2 = O_p\left(\frac{d_0\log\left(\frac{d}{d_0}\right)}{n}\log\left(d_0\log\left(\frac{d}{d_0}\right)\right)\right) \tag{B.84}$$

□

#### B.0.2.4 Proof of Theorem B.6

Out of convenience we re-state the theorem: Let $\mathcal{I}$ be a subset of $S^{d-1}$ and define

$$\hat{\theta}_{\mathcal{I}} := \arg\min_{\theta\in\mathcal{I}} \mathbb{M}^*_{n,d}(\theta). \tag{B.85}$$

For any $s > 0$, with probability greater than $1 - e^{-s}$:

$$\left|R^*_d(\hat{\theta}_{\mathcal{I}}) - \min_{\theta\in\mathcal{I}} R^*_d(\theta)\right| \;\leq\; 2K\sqrt{\frac{V_n}{n}} + 2\sqrt{\frac{1}{n}} + 2\sqrt{\frac{2s}{n}} + \frac{4}{n}$$

$$\left|R_d^*(\hat{\theta}_{\mathcal{I}}) - \mathbb{M}_{n,d}^*(\hat{\theta}_{\mathcal{I}})\right| \leq K\sqrt{\frac{V_n}{n}} + \sqrt{\frac{1}{n}} + \sqrt{\frac{2s}{n}} + \frac{2}{n} \tag{B.86}$$

where $V_n$ is the VC dimension of $\mathcal{I}$ and $K$ is a universal constant.

Overall the proof will proceed very similar to the proof of Theorem 9.1 in *Giraud* (2014).

*Proof.* Define $\hat{\Delta}_{n,d} := \sup_{\Delta \in \mathcal{I}} \left|\mathbb{M}_{n,d}^*(\theta) - R_d^*(\theta)\right|$. We have

$$\begin{aligned} \left|R_d^*(\hat{\theta}_{\mathcal{I}}) - \min_{\theta \in \mathcal{I}} R_d^*(\theta)\right| &\leq 2\hat{\Delta}_{n,d}(\mathcal{I}) \\ \left|R_d^*(\hat{\theta}_{\mathcal{I}}) - \mathbb{M}_{n,d}^*(\hat{\theta}_{\mathcal{I}})\right| &\leq \hat{\Delta}_{n,d}(\mathcal{I}) \end{aligned} \tag{B.87}$$

This is not hard to see. By definition, $\mathbb{M}_{n,d}^*(\hat{\theta}_{\mathcal{I}}) \leq \mathbb{M}_{n,d}^*(\theta)$ for all $\theta \in \mathcal{I}$, which allows us to derive

$$\begin{aligned} \left|R_d^*(\hat{\theta}_{\mathcal{I}}) - R_d^*(\theta)\right| &= \left|R_d^*(\hat{\theta}_{\mathcal{I}}) - \mathbb{M}_{n,d}^*(\hat{\theta}_{\mathcal{I}}) + \mathbb{M}_{n,d}^*(\hat{\theta}_{\mathcal{I}}) - R_d^*(\theta)\right| \\ &\leq \left|R_d^*(\hat{\theta}_{\mathcal{I}}) - \mathbb{M}_{n,d}^*(\hat{\theta}_{\mathcal{I}})\right| + \left|\mathbb{M}_{n,d}^*(\theta) - R_d^*(\theta)\right| \\ &\leq 2\sup_{\theta \in \mathcal{I}} \left|\mathbb{M}_{n,d}^*(\theta) - R_d^*(\theta)\right| \end{aligned} \tag{B.88}$$

With this bound, it will be sufficient to show that $\hat{\Delta}_{n,d}$ is no greater than the right hand sides of (B.57). This will be done in two parts, with the first part showing $\hat{\Delta}_{n,d}$ does not exceed its expected value with an exponentially decreasing probability, and a second part deriving a bound for $\mathbb{E}\hat{\Delta}_{n,d}$.

**Probability Bound on $\hat{\Delta}_{n,d}$**: Suppose $s > 0$, then with probability at least $1 - e^{-s}$, we have

$$\hat{\Delta}_{n,d}(\mathcal{I}) \leq \mathbb{E}[\hat{\Delta}_{n,d}(\mathcal{I})] + \sqrt{\frac{2s}{n}} + \frac{2}{n} \tag{B.89}$$

To show this first note that a bound the value of $\hat{\Delta}_{n,d}(\theta)$ can be obtained by obtaining a bound on each of $\sup_{\theta\in\mathcal{I}} \left|\mathbb{M}^*_{n,d}(\theta) - \mathbb{E}[\mathbb{M}^*_{n,d}(\theta)]\right|$ and $\sup_{\theta\in\mathcal{I}} |\mathbb{E}[\mathbb{M}^*_{n,d}] - R^*_d(\theta)|$.

To obtain a bound for $\sup_{\theta\in\mathcal{I}} |\mathbb{E}[\mathbb{M}^*_{n,d}] - R^*_d(\theta)|$, note that for any fixed $\theta \in \mathcal{I}$, we have

$$\begin{aligned}
& \left|R^*_d(\theta) - \mathbb{E}[\mathbb{M}^*_{n,d}(\theta)]\right| \\
= \; & \left|\mathbb{E}\left[L^*_{n,d}(\theta) - \mathbb{M}^*_{n,d}(\theta)\right]\right| \\
= \; & \left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\bar{Y}-\gamma)\left(\mathbb{1}(X_i^T\theta \le 0) - \mathbb{1}(X_i^T\theta_0 \le 0)\right)\right]\right| \\
= \; & \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\bar{Y}-\gamma)\left(\mathbb{1}(X_i^T\theta \le 0) - \mathbb{1}(X_i^T\theta_0 \le 0)\right)\right]\right|. \qquad \text{(B.90)}
\end{aligned}$$

For each $j = 1, \ldots, n$,

$$\begin{aligned}
& \left|\mathbb{E}\left[(\bar{Y}-\gamma)\left(\mathbb{1}(X_j^T\theta \le 0) - \mathbb{1}(X_j^T\theta_0 \le 0)\right)\right]\right| \\
= \; & \left|\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}(Y_k-\gamma)\left(\mathbb{1}(X_j^T\theta \le 0) - \mathbb{1}(X_j^T\theta_0 \le 0)\right)\right]\right| \\
= \; & \left|\frac{1}{n}\sum_{k\neq j}\mathbb{E}\left[(Y_k-\gamma)\right]\mathbb{E}\left[\mathbb{1}(X_j^T\theta \le 0) - \mathbb{1}(X_j^T\theta_0 \le 0)\right]\right| \\
& + \frac{1}{n}\mathbb{E}\left[\left|(Y_j-\gamma)\left(\mathbb{1}(X_j^T\theta \le 0) - \mathbb{1}(X_j^T\theta_0 \le 0)\right)\right|\right] \\
\le \; & \frac{1}{n}, \qquad \text{(B.91)}
\end{aligned}$$

and therefore, $0 \le \left|R^*_d(\theta) - \mathbb{E}[\mathbb{M}^*_{n,d}(\theta)]\right| \le \frac{1}{n}$ for all $\theta \in S^{d-1}$.

As for bounding $\left|\mathbb{M}^*_{n,d}(\theta) - \mathbb{E}[\mathbb{M}^*_{n,d}(\theta)]\right|$, define the function

$$F\Big((X_1,Y_1),\ldots,(X_n,Y_n)\Big) := \sup_{\theta\in\mathcal{I}}\left|\mathbb{M}^*_{n,d}(\theta) - R^*_d(\theta)\right|. \qquad \text{(B.92)}$$

Now suppose that for some $j \in \{1, \ldots, n\}$, we have $(X_j, Y_j) \neq (X'_j, Y'_j)$. Then we would have

$$\left(\frac{1}{n}\sum_{k\neq j}^{n} Y_k\right) + \frac{Y'_j}{n} = \bar{Y} + \frac{Y'_j - Y_j}{n} \tag{B.93}$$

and hence

$$\begin{aligned}
& F\Big((X_1,Y_1),\ldots,(X_{j-1},Y_{j-1}),(X'_j,Y'_j),(X_{j+1},Y_{j+1}),\ldots,(X_n,Y_n)\Big) \\
& -F\Big((X_1,Y_1),\ldots,(X_n,Y_n)\Big) \\
= \; & \sup_{\theta\in\mathcal{I}} \Bigg| \frac{1}{n}\sum_{k\neq j}\left(Y_k - \bar{Y} - \frac{Y'_j - Y_j}{n}\right)\big(\mathbb{1}(X_k^T\theta \leq 0) - \mathbb{1}(X_k^T\theta_0 \leq 0)\big) \\
& + \frac{1}{n}\left(Y'_j - \bar{Y} - \frac{Y'_j - Y_j}{n}\right)\big(\mathbb{1}(X_j^T\theta \leq 0) - \mathbb{1}(X_j^T\theta_0 \leq 0)\big) - R_d^*(\theta) \Bigg| \\
& - \sup_{\theta\in\mathcal{I}} \left| \frac{1}{n}\sum_{k=1}^{n}\big(Y_k - \bar{Y}\big)\big(\mathbb{1}(X_k^T\theta \leq 0) - \mathbb{1}(X_k^T\theta_0 \leq 0)\big) - R_d^*(\theta)\right| \\
\leq \; & \sup_{\theta\in\mathcal{I}} \Bigg| \frac{1}{n}\sum_{k\neq j}\left(Y_k - \bar{Y} - \frac{Y'_j - Y_j}{n}\right)\big(\mathbb{1}(X_k^T\theta \leq 0) - \mathbb{1}(X_k^T\theta_0 \leq 0)\big) \\
& + \frac{1}{n}\left(Y'_j - \bar{Y} - \frac{Y'_j - Y_j}{n}\right)\big(\mathbb{1}(X_j^T\theta \leq 0) - \mathbb{1}(X_j^T\theta_0 \leq 0)\big) - R_d^*(\theta) \\
& - \left(\frac{1}{n}\sum_{k=1}^{n}\big(Y_k - \bar{Y}\big)\big(\mathbb{1}(X_k^T\theta \leq 0) - \mathbb{1}(X_k^T\theta_0 \leq 0)\big) - R_d^*(\theta)\right) \Bigg| \\
= \; & \sup_{\theta\in\mathcal{I}} \left| \frac{1}{n}\sum_{k\neq j}\left(\frac{Y_j - Y'_j}{n}\right)\big(\mathbb{1}(X_k^T\theta \leq 0) - \mathbb{1}(X_k^T\theta_0 \leq 0)\big) \right. \\
& \left. + \frac{n-1}{n^2}(Y'_j - Y_j)\big(\mathbb{1}(X_j^T\theta \leq 0) - \mathbb{1}(X_j^T\theta_0 \leq 0)\big) \right| \\
\leq \; & \frac{1}{n}\cdot(n-1)\cdot\frac{1}{n} + \frac{n-1}{n^2} \\
< \; & \frac{2}{n}.
\end{aligned} \tag{B.94}$$

Using McDiarmid's inequality, this leads to the deduction that

$$\mathbb{P}\left[\sup_{\theta\in\mathcal{I}}\left|\mathbb{M}_{n,d}^*(\theta) - \mathbb{E}[\mathbb{M}_{n,d}^*(\theta)]\right| > \mathbb{E}\left[\sup_{\theta\in\mathcal{I}}\left|\mathbb{M}_{n,d}^*(\theta) - \mathbb{E}[\mathbb{M}_{n,d}^*(\theta)]\right|\right] + t\right] \leq \exp\left[-\frac{nt^2}{2}\right]$$

(B.95)

with probability greater than $1 - e^{-t}$.

In conclusion, since $\hat{\Delta}_{n,d}(\mathcal{I}) = \sup_{\theta\in\mathcal{I}} \left|\mathbb{M}^*_{n,d}(\theta) - \mathbb{E}\mathbb{M}^*_{n,d}(\theta) + \mathbb{E}\mathbb{M}^*_{n,d}(\theta) - R^*_d(\theta)\right|$,

$$\begin{aligned}
& \mathbb{P}\left[\hat{\Delta}_{n,d}(\mathcal{I}) > \mathbb{E}[\hat{\Delta}_{n,d}(\mathcal{I})] + \sqrt{\frac{2s}{n}} + \frac{2}{n}\right] \\
\leq\ & \mathbb{P}\left[\sup_{\theta\in\mathcal{I}} \left|\mathbb{M}^*_{n,d}(\theta) - \mathbb{E}\mathbb{M}^*_{n,d}(\theta)\right| + \left(\sup_{\theta\in\mathcal{I}} \left|\mathbb{E}\mathbb{M}^*_{n,d}(\theta) - R^*_d(\theta)\right| - \frac{1}{n}\right) > \mathbb{E}[\hat{\Delta}_{n,d}(\mathcal{I})] + \sqrt{\frac{2s}{n}} + \frac{1}{n}\right] \\
\leq\ & \mathbb{P}\left[\sup_{\theta\in\mathcal{I}} \left|\mathbb{M}^*_{n,d}(\theta) - \mathbb{E}\mathbb{M}^*_{n,d}(\theta)\right| > \mathbb{E}[\hat{\Delta}_{n,d}(\mathcal{I})] + \sqrt{\frac{2s}{n}} + \frac{1}{n}\right] \\
\leq\ & \mathbb{P}\left[\sup_{\theta\in\mathcal{I}} \left|\mathbb{M}^*_{n,d}(\theta) - \mathbb{E}\mathbb{M}^*_{n,d}(\theta)\right| > \mathbb{E}\left[\sup_{\theta\in\mathcal{I}} \left|\mathbb{M}^*_{n,d}(\theta) - \mathbb{E}\mathbb{M}^*_{n,d}(\theta)\right|\right] + \sqrt{\frac{2s}{n}}\right. \\
& \left. + \left(\frac{1}{n} - \sup_{\theta\in\mathcal{I}} \left|\mathbb{E}\mathbb{M}^*_{n,d}(\theta) - R^*_d(\theta)\right|\right)\right] \\
\leq\ & \mathbb{P}\left[\sup_{\theta\in\mathcal{I}} \left|\mathbb{M}^*_{n,d}(\theta) - \mathbb{E}\mathbb{M}^*_{n,d}(\theta)\right| > \mathbb{E}\left[\sup_{\theta\in\mathcal{I}} \left|\mathbb{M}^*_{n,d}(\theta) - \mathbb{E}\mathbb{M}^*_{n,d}(\theta)\right|\right] + \sqrt{\frac{2s}{n}}\right] \\
\leq\ & \exp\left[-s\right]
\end{aligned} \tag{B.96}$$

**Bound on** $\mathbb{E}[\hat{\Delta}_{n,d}(\mathcal{I})]$: We next show that $\mathbb{E}[\hat{\Delta}_{n,d}(\mathcal{I})] \leq K\sqrt{\frac{V_n}{n}} + \sqrt{\frac{1}{n}}$, where $V_n$ is the VC dimension of $\mathcal{I}$ and $K$ is a universal constant. First we separate

$$\begin{aligned}
\mathbb{E}[\hat{\Delta}_{n,d}(\mathcal{I})] &= \mathbb{E}\left[\sup_{\theta\in\mathcal{I}} \left|\mathbb{M}^*_{n,d}(\theta) - L^*_{n,d}(\theta) + L^*_{n,d}(\theta) - R^*_d(\theta)\right|\right] \\
&\leq \mathbb{E}\left[\sup_{\theta\in\mathcal{I}} \left|\mathbb{M}^*_{n,d}(\theta) - L^*_{n,d}(\theta)\right|\right] + \mathbb{E}\left[\sup_{\theta\in\mathcal{I}} \left|L^*_{n,d}(\theta) - R^*_d(\theta)\right|\right]
\end{aligned} \tag{B.97}$$

To bound the first term, note that for every possible $\theta$,

$$\begin{aligned}
\left|\mathbb{M}^*_{n,d}(\theta) - L^*_{n,d}(\theta)\right| &= \left|(\bar{Y} - \gamma)\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}(X_i^T\theta \leq 0) - \mathbb{1}(X_i^T\theta_0 \leq 0)\right)\right| \\
&\leq |\bar{Y} - \gamma|
\end{aligned} \tag{B.98}$$

and hence

$$\begin{aligned}\mathbb{E}\left[\sup_{\theta\in\mathcal{I}}\left|\mathbb{M}^*_{n,d}(\theta)-L^*_{n,d}(\theta)\right|\right] &\leq \mathbb{E}[|\bar{Y}-\gamma|] \\ &\leq \sqrt{\mathbb{E}[(\bar{Y}-\gamma)^2]} \\ &= \sqrt{\frac{\gamma(1-\gamma)}{n}}. \end{aligned} \tag{B.99}$$

As for the second term,

$$\begin{aligned}&\mathbb{E}\left[\sup_{\theta\in\mathcal{I}}\left|L^*_{n,d}(\theta)-R^*_d(\theta)\right|\right] \\ =\ &\mathbb{E}\left[\sup_{\theta\in\mathcal{I}}\left|(\mathbb{P}_n-P)\Big((Y-\gamma)\left(\mathbb{1}(X^T\theta\leq 0)-\mathbb{1}(X^T\theta_0\leq 0)\right)\Big)\right|\right] \\ \leq\ &2\mathbb{E}\left[\sup_{\theta\in\mathcal{I}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\Big((Y_i-\gamma)\left(\mathbb{1}(X_i^T\theta\leq 0)-\mathbb{1}(X_i^T\theta_0\leq 0)\right)\Big)\right|\right] \\ &\text{where } \varepsilon_i\text{'s are iid Rademacher variables} \\ \leq\ &K_1\frac{1}{\sqrt{n}}\int_0^\infty\sqrt{\log N(\eta,\mathcal{F}_\mathcal{I},L_2(\mathbb{P}_n))}\,d\eta \end{aligned} \tag{B.100}$$

for some universal constant $K$, and where $\mathcal{F}_\mathcal{I}$ is defined as the set of functions $\{(Y-\gamma)\left(\mathbb{1}(X^T\theta\leq 0)-\mathbb{1}(X^T\theta_0\leq 0)\right) : \theta\in\mathcal{I}\}$. Because this class of functions has the envelope function $F(X,Y)=1$, this means that for all $\eta\leq 1$,

$$\begin{aligned}\log N(\eta,\mathcal{F}_\mathcal{I},L_2(\mathbb{P}_n)) &= \log N(\eta\|F\|_{\mathbb{P}_n},\mathcal{F}_\mathcal{I},L_2(\mathbb{P}_n)) \\ &\leq K_2V_n(\log(1/\eta)+K_3) \end{aligned} \tag{B.101}$$

for some universal constant $K_2, K_3$, and where $V_n$ is the VC dimension of $\mathcal{I}$. Therefore

$$\begin{aligned}&\mathbb{E}\left[\sup_{\theta\in\mathcal{I}}\left|L^*_{n,d}(\theta)-R^*_d(\theta)\right|\right] \\ \leq\ &K_1\sqrt{\frac{V_n}{n}}\int_0^1\sqrt{K_2\log(1/\eta)+K_2K_3}\,d\eta\end{aligned}$$

$$\leq \quad K\sqrt{\frac{V_n}{n}} \tag{B.102}$$

for some universal constant $K$. □

### B.0.3 Proof of Theorem III.9

To derive the consistency of the least squares estimator, we will utilize empirical process theory for the class of functions $\mathcal{F}_n := \{f_\theta : \theta \in S^{d_n-1}\}$, where $S^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ and

$$f_\theta(y, x) = \left(Y - \frac{\mu}{2}\right)\left(1(X^T\theta > 0) - 1(X^T\theta_0 > 0)\right) \tag{B.103}$$

A few observations can be made regarding the class of functions $\mathcal{F}_n := \{f_\theta : \theta \in \mathbb{R}^d, \|\theta\|_2 = 1\}$. The first is that it is a VC class of functions

**Lemma 25.** *For any $n$, $\mathcal{F}_n$ is a VC class of functions with VC class at most $d + 2$.*

*Proof.* The class of functions $\{g(x) = x^T\theta : \theta \in \mathbb{R}^d\}$ is a $d$ dimensional vector space of functions with VC dimension at most $d + 2$ (see Lemma 2.6.15 of VDVW), and thus the same holds for its subset $\mathcal{G}_d := \{g(x) = x^T\theta : \theta \in \mathbb{R}^d, \|\theta\| = 1\}$. Since the indicator functions $\phi(y) = 1(y > 0)$ is monotone, $\phi(\mathcal{G}_d)$ is a VC class with VC dimension at most $d + 2$ (see Lemma 2.6.18 of VDVW), and multiplying by the real valued function $h(y) = y - \frac{\mu}{2}$ makes $\mathcal{F}_d = \phi(\mathcal{G}_d) \cdot h$ have the same property. □

The class $\mathcal{F}_n$ has the envelope function $F_n(Y, X) := |Y - \mu 1(X^T\theta_0 > 0)| + \frac{3\mu}{2} = |\epsilon| + \frac{3\mu}{2}$, which is square integrable over the probability space as $PF_n^2 \leq \sigma^2 + \mu + 2 \cdot \frac{\mu^2}{4}$. Following Theorem 6.2 in Wellner's notes yields the following bound:

**Lemma 26.** *For any $\epsilon > 0$ there exists a constant $M_\epsilon$ such that*

$$E^*\|\mathbb{P}_n - P\|_{\mathcal{F}_n} \leq M_\epsilon E^* \sqrt{\frac{1 + \log N(\epsilon, \mathcal{F}_n, L_2(\mathbb{P}_n))}{n}} + \epsilon \tag{B.104}$$

*for all $n$.*

From here a Glivenko-Cantelli-type result can be established for the class of functions $\mathcal{F}_n$:

**Lemma 27.** *Under Assumption B1 and B2, $E^*\|\mathbb{P}_n - P\|_{\mathcal{F}_n} \to 0$.*

*Proof.* From Theorem 2.6.7 of VDVW,

$$N(\epsilon, \mathcal{F}_n, L_2(\mathbb{P}_n)) \leq KV(\mathcal{F}_n)\left(\frac{16e\|F\|_{\mathbb{P}_n,2}}{\epsilon}\right)^{2V(\mathcal{F}_n)-2} \tag{B.105}$$

for any fixed $\epsilon > 0$. Since the VC dimension is at most $p+2$, this means

$$\sqrt{\frac{1+\log(N(\epsilon, \mathcal{F}_n, L_2(\mathbb{P}_n)))}{n}} \leq \sqrt{\frac{1+\log(n+2)+(2d_n-1)(\log(\|F_n\|_{\mathbb{P}_n,2})-\log(\epsilon)+C_\epsilon)}{n}} \tag{B.106}$$

for some constant $C_\epsilon$. This gives the following bound when plugged in for (26):

$$\begin{aligned}
& E^*\|\mathbb{P}_n - P\|_{\mathcal{F}_n} \\
\leq\ & M_\epsilon E^* \sqrt{\frac{1+\log N(\epsilon, \mathcal{F}_n, L_2(\mathbb{P}_n))}{n}} + \epsilon \\
\leq\ & M_\epsilon \left( \sqrt{\frac{1+\log(d+2)+2d(C_\epsilon - \log(\epsilon))}{n}} + E^* \sqrt{\frac{2d_n(\log(\|F\|_{\mathbb{P}_n,2}) \vee 0)}{n}} \right) + \epsilon \\
\leq\ & M_\epsilon \left( \sqrt{\frac{1+\log(d+2)+2d(C_\epsilon - \log(\epsilon))}{n}} + E^* \sqrt{\frac{2d}{n}} \|F\|_{\mathbb{P}_n,2} \right) + \epsilon \\
& \text{since for all positive } x, \sqrt{\log(x) \vee 0} \leq x \\
\to\ & \epsilon
\end{aligned} \tag{B.107}$$

Since $\epsilon$ is arbitrary, $E^*\|\mathbb{P}_n - P\|_{\mathcal{F}_n} \to 0$. □

To establish the consistency of $\hat{\theta}$ for $\theta$, it needs to be established that $Pf_\theta(Y,X)$ is well separated from $Pf_{\theta_0}(Y,X)$ when $\theta$ is well separated from $\theta_0$, which is assured by Assumption B2.

**Lemma 28.** *For all sufficiently small $\eta > 0$,*

$$\inf_{\|\theta-\theta_0\|_2\geq\eta} \mathbb{E}\left|\left(Y-\frac{\mu}{2}\right)\left(1(X^T\theta>0)-1(X^T\theta_0>0)\right)\right| \geq C\eta \tag{B.108}$$

*for some constant $C$ which does not depend on $n$.*

*Proof.* For any $\theta$ with norm 1,

$$\begin{aligned} & \mathbb{E}\left|\left(Y-\frac{\mu}{2}\right)\left(1(X^T\theta>0)-1(X^T\theta_0>0)\right)\right| \\ = \; & \mathbb{E}_X\mathbb{E}_\varepsilon\left|\mu 1(X^T\theta_0>0)+\varepsilon-\frac{\mu}{2}\right|\cdot\left|1(X^T\theta>0)-1(X^T\theta_0>0))\right| \end{aligned} \tag{B.109}$$

In the case where $|\varepsilon| = \frac{\mu}{2}$ with probability strictly less than 1, the above equals

$$\begin{aligned} & \mathbb{E}_X\mathbb{E}_\epsilon\left|\frac{\mu}{2}-|\varepsilon|\right|\cdot\left|1(X^T\theta>0)-1(X^T\theta_0>0))\right| \\ = \; & \mathbb{E}_\epsilon\left|\frac{\mu}{2}-|\varepsilon|\right|\cdot\mathbb{E}_X\left|1(X^T\theta>0)-1(X^T\theta_0>0))\right| \\ \geq \; & a^-\|\theta-\theta_0\|_2 \end{aligned} \tag{B.110}$$

On the other hand, if $|\varepsilon| = \frac{\mu}{2}$ with probability 1, ie $P\left(\varepsilon=\frac{\mu}{2}\right) = P\left(\varepsilon=\frac{\mu}{2}\right) = 1/2$, then regardless of the value of $X^T\theta_0$,

$$P\left(\left|\mu 1(X^T\theta_0>0)+\varepsilon-\frac{\mu}{2}\right|=0\middle|X^T\theta_0\right) = P\left(\left|\mu 1(X^T\theta_0>0)+\varepsilon-\frac{\mu}{2}\right|=\mu\middle|X^T\theta_0\right) = \frac{1}{2} \tag{B.111}$$

Hence $\left|\mu 1(X^T\theta_0>0)+\varepsilon-\frac{\mu}{2}\right|$ is independent of $X$ and

$$\begin{aligned} & \mathbb{E}_X\mathbb{E}_\varepsilon\left|\mu 1(X^T\theta_0>0)+\varepsilon-\frac{\mu}{2}\right|\cdot\left|1(X^T\theta>0)-1(X^T\theta_0>0))\right| \\ = \; & \mathbb{E}_X\mathbb{E}_\epsilon\left|\frac{\mu}{2}-|\varepsilon|\right|\cdot\left|1(X^T\theta>0)-1(X^T\theta_0>0))\right| \\ = \; & \mathbb{E}_\epsilon\left|\frac{\mu}{2}-|\varepsilon|\right|\cdot\mathbb{E}_X\left|1(X^T\theta>0)-1(X^T\theta_0>0))\right| \\ \geq \; & a^-\|\theta-\theta_0\|_2 \end{aligned} \tag{B.112}$$

Therefore,

$$\begin{aligned} & \inf_{\|\theta-\theta_0\|\geq\eta} \mathbb{E}\left|\left(Y-\frac{\mu}{2}\right)\left(1(X^T\theta>0)-1(X^T\theta_0>0)\right)\right| \\ = & \inf_{\|\theta-\theta_0\|\geq\eta} \mathbb{E}\left|\frac{\mu}{2}(1(X^T\theta>0)-1(X^T\theta_0>0))\right| \\ \geq & \inf_{\|\theta-\theta_0\|\geq\eta} C\|\theta-\theta_0\|_2 \\ \geq & C\eta \end{aligned} \tag{B.113}$$

for some constant $C$ not dependent on $n$. □

Combining the two lemmas above can demonstrate the consistency of $\hat{\theta}^{(sq)}$.

**Theorem B.7.** $\|\hat{\theta}^{(sq)}-\theta_0\|_2 \to 0$ *in probability.*

*Proof.* Due the previous lemm, for any $\epsilon>0$, and some constant $C>0$,

$$\begin{aligned} & \|\hat{\theta}^{(sq)}-\theta_0\|_2 \geq \epsilon \\ \to & Pf_{\theta_0}-Pf_{\hat{\theta}^{(sq)}} \geq C\epsilon \\ \to & Pf_{\hat{\theta}^{(sq)}}-Pf_{\theta_0}-(\mathbb{P}_n f_{\hat{\theta}^{(sq)}}-\mathbb{P}_n f_{\theta_0}) \leq -C\epsilon \\ \to & \sup_{\|\theta\|_2=1} |Pf_\theta-\mathbb{P}_n f_\theta| \geq C\epsilon/2 \end{aligned} \tag{B.114}$$

Therefore,

$$\begin{aligned} & P\left[\|\hat{\theta}^{(sq)}-\theta_0\| \geq \epsilon\right] \\ \leq & P\left[\sup_{\|\theta\|_2=1} |Pf_\theta-\mathbb{P}_n f_\theta| \geq C\epsilon/2\right] \\ \leq & P\left[|P-\mathbb{P}_n|_{\mathcal{F}_n} \geq C\epsilon/2\right] \\ \to & 0 \end{aligned} \tag{B.115}$$

□

### B.0.4 Proof of Theorem 3.6

To prove the rate of convergence, we will again utilize Theorem B.2. For the model under consideration, the notation of this theorem would correspond to

$$\begin{aligned}
\mathbb{M}_n(\theta) &:= \mathbb{P}_n f_\theta(X,Y) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(Y-\frac{\mu}{2}\right)\left[1(X^T\theta>0)-1(X^T\theta_0>0)\right] \\
M_n(\theta) &:= Ef_\theta(X,Y) \\
&= E\left[\left(Y-\frac{\mu}{2}\right)\left[1(X^T\theta>0)-1(X^T\theta_0>0)\right]\right] \qquad \text{(B.116)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
& M_n(\theta)-M_n(\theta_0) \\
= \; & E_X E_\varepsilon\left[\left(\mu 1(X^T\theta_0>0)+\varepsilon-\frac{\mu}{2}\right)\left[1(X^T\theta>0)-1(X^T\theta_0>0)\right]\right] \\
= \; & -\frac{\mu}{2}E_X\left[1(X^T\theta>0, X^T\theta_0\le 0)+1(X^T\le 0, X^T\theta_0>0)\right] \\
= \; & -\frac{\mu}{2}E\left|1(X^T\theta>0)-1(X^T\theta_0>0)\right| \qquad \text{(B.117)}
\end{aligned}$$

Assumption B1 means that for all $\theta$ sufficiently close to $\theta_0$, the last line $\lesssim -\|\theta-\theta_0\|_2$. Hence the relevant distance function here is $\rho_n(\theta,\theta_0) := \sqrt{\|\theta-\theta_0\|_2}$. Consequently, define the set of functions $\mathcal{F}_{n,\delta} := \{f_\theta \in \mathcal{F}_n : \rho_n^2(\theta_0,\theta)\le\delta\}$.

We will bound the modulus of continuity over the set $\mathcal{F}_{n,\delta}$. Let

$$\begin{aligned}
\mathbb{G}_n(\theta) &= \sqrt{n}(f_\theta - Ef_\theta) \\
&= \sqrt{n}\Big[\left(Y-\frac{\mu}{2}\right)\left[1(X^T\theta>0)-1(X^T\theta_0>0)\right] \\
&\quad -E\left(Y-\frac{\mu}{2}\right)\left[1(X^T\theta>0)-1(X^T\theta_0>0)\right]\Big]. \qquad \text{(B.118)}
\end{aligned}$$

By Theorem 2.14.1 of VDVW,

$$E[\|\mathbb{G}_n\|_{\mathcal{F}_{n,\delta}}] \lesssim E[J(\theta_n, \mathcal{F}_{n,\delta})\|F_n\|_n] \tag{B.119}$$

where

$$J(\nu_n, \mathcal{F}_{p,\delta}) = \sup_Q \int_0^{\nu_n} \sqrt{1 + \log(N(\epsilon\|F\|_{Q,2}, \mathcal{F}_{n,\delta}, L_2(Q)))}\, d\epsilon$$

$$\nu_n = \frac{\sup_{f\in\mathcal{F}_{n,\delta}} \sqrt{\frac{1}{n}\sum_{i=1}^n f^2(X_i, Y_i)}}{\sqrt{\frac{1}{n}\sum_{i=1}^n F_n(X_i, Y_i)^2}} \tag{B.120}$$

The expression in (B.119) will be easier to work with on an event space where $\|F_n\|_n$ is bounded above by a constant. Explicitly, the expression equals

$$\begin{aligned} \|F_n\|_n &= \sqrt{\frac{1}{n}\sum_{i=1}^n \left(|Y_i - 1(X_i^T\theta_0 > 0)| + \frac{\mu}{2}\right)} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^n \left(|\varepsilon_i| + \frac{\mu}{2}\right)}. \end{aligned} \tag{B.121}$$

For any positive constant $A$ such that $A > E\left[|\varepsilon| + \frac{\mu}{2}\right]$, the probability that $\|F\|_n$ exceeds $\sqrt{A}$ goes to 0 because although the error terms $\epsilon_i = \epsilon_{i,n}$ technically form a triangular array of random variables, they are still iid random variables with the same distribution for each $n$, permitting the use of Chebychev's inequality:

$$\begin{aligned} P\left[\|F_n\|_n > \sqrt{A}\right] &= P\left[\|F_n\|_n^2 > A\right] \\ &= P\left[\frac{1}{n}\sum_{i=1}^n \left(|\varepsilon_i| + |\frac{\mu}{2}\right) - E\left[|\varepsilon| + \frac{\mu}{2}\right] > A - E\left[|\varepsilon| + \frac{\mu}{2}\right]\right] \\ &\leq \left(\frac{A - E\left[|\varepsilon| + |\frac{\mu}{2}\right]}{\sqrt{\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n \left(|\varepsilon_i| + \frac{\mu}{2}\right)\right)}}\right)^{-2} \end{aligned}$$

$$
\begin{aligned}
&= \left( \frac{A - E\left[|\varepsilon| + |\frac{\mu}{2}\right]}{\sqrt{\mathrm{Var}\left(\left(|\varepsilon| + \frac{\mu}{2}\right)\right)}/\sqrt{n}} \right)^{-2} \\
&\to 0.
\end{aligned} \tag{B.122}
$$

Hence, take $\Omega_n := \{\|F_n\|_n \leq \sqrt{A}\}$ for any fixed constant $A$ with $A > E\left[|\varepsilon| + \frac{\mu}{2}\right]$, then $P(\Omega_n) \to 1$, and using a modification of Theorem 2.14.1 of VDVW,

$$
\begin{aligned}
E[\|\mathbb{G}_n 1(\Omega_n)\|_{\mathcal{F}_{n,\delta}}] &\lesssim E[J(\nu_n, \mathcal{F}_{n,\delta})\|F\|_n 1(\Omega_n)] \\
&\leq AE[J(\nu_n, \mathcal{F}_{n,\delta})]
\end{aligned} \tag{B.123}
$$

From here it suffices to bound the value of $E[J(\nu_n, \mathcal{F}_{n,\delta})]$. Since $\mathcal{F}_{n,\delta}$ is a subset of $\mathcal{F}_n$, and hence has a VC dimension of some constant times $d$. This gives a bound on the covering numbers:

$$
\log N(\eta, \mathcal{F}_n, Q) \leq -Cd_n \log(\eta) \tag{B.124}
$$

for some constant $C$, and hence

$$
\begin{aligned}
J(\nu_n, \mathcal{F}_{n,\delta}) &\lesssim \int_0^{\nu_n} \sqrt{d \log(1/\eta)}\, d\eta \\
&= \sqrt{d} \int_{-\log \eta}^{\infty} \sqrt{z} e^{-z}\, dz \\
&= \sqrt{d}\Gamma(3/2, -\log(\nu_n)) \\
&= \sqrt{d}[\Gamma(1/2, -\log(\nu_n)) + \nu_n \sqrt{-\log(\nu_n)}] \\
&= \sqrt{d}\left[\sqrt{\pi}\mathrm{erfc}(\sqrt{-\log(\nu_n)}) + \nu_n \sqrt{-\log \nu_n}\right] \\
&\leq \sqrt{d}[\sqrt{\pi}\nu_n + \nu_n \sqrt{-\log \nu_n}] \qquad \text{since } \mathrm{erfc}(x) \leq e^{-x^2}
\end{aligned} \tag{B.125}
$$

For values of $\nu_n$ around a small neighborhood of 0, this is bounded above by a constant (not dependent on $n$) times $\sqrt{d}[\nu_n\sqrt{-\log\nu_n}]$. The function $x \to x\sqrt{-\log(x)}$ has a strictly negative second derivative for $x \in (0,1)$, hence Jensen's inequality gives

$$EJ(\nu_n, \mathcal{F}_{n,\delta}) \lesssim E[\sqrt{d}\nu_n\sqrt{-\log(\nu_n)}] \lesssim \sqrt{d}E(\nu_n)\sqrt{-\log E(\nu_n)} \tag{B.126}$$

From here it suffices to bound the value of $E(\nu_n)$, from above. Note that since $F(x,y) = (y - 1(x^T\theta_0 > 0)) + \frac{\mu}{2} \geq \frac{\mu}{2}$, this means

$$\nu_n \leq \frac{2}{\mu}\sup_{f\in\mathcal{F}_{n,\delta}}\sqrt{\frac{1}{n}\sum_{i=1}^n f^2(X_i, Y_i)} \tag{B.127}$$

Therefore,

$$\begin{aligned} E(\nu_n) &= E\left(\frac{\sup_{f\in\mathcal{F}_{n,\delta}}\frac{1}{n}\sum_{i=1}^n f^2(X_i,Y_i)}{\|F\|_{\mathbb{P}_n}}\right) \\ &\leq \frac{2}{\mu}E\left(\sup_{f\in\mathcal{F}_{n,\delta}}\frac{1}{n}\sum_{i=1}^n f^2(X_i,Y_i)\right) \\ &\leq \frac{2}{\mu}\sqrt{E\left(\sup_{f\in\mathcal{F}_{n,\delta}}\sum_{i=1}^n\left(Y_i - \frac{\mu}{2}\right)^2|1(X_i^T\theta > 0) - 1(X_i^T\theta_0 > 0)|\right)} \\ &= \frac{2}{\mu}\left[E\left(\sup_{g\in\mathcal{G}_{n,\delta}}\mathbb{P}_n(g)\right)\right]^{1/2} \end{aligned} \tag{B.128}$$

where $\mathcal{G}_{n,\delta} := \{f^2 : f \in \mathcal{F}_{n,\delta}\}$. To bound the above, one way is to separate the expectation into two terms:

$$E\left[\sup_{g\in\mathcal{G}_{n,\delta}}\mathbb{P}_n(g)\right] \leq E\left[\sup_{g\in\mathcal{G}_{n,\delta}}|\mathbb{P}_n - P|(g)\right] + \sup_{g\in\mathcal{G}_{n,\delta}}P(g) \tag{B.129}$$

The first term can be bounded as follows:

$$
\begin{aligned}
& E\left[\sup_{g\in\mathcal{G}_{n,\delta}} |\mathbb{P}_n - P|(g)\right] \\
= \;& \frac{1}{\sqrt{n}} E[\|\mathbb{G}_n\|_{\mathcal{G}_{n,\delta}}] \\
\lesssim \;& \frac{1}{\sqrt{n}} E_X\left[E_Z\left[\left\|\frac{1}{\sqrt{n}}\sum_{i=1} n\epsilon_i g(X_i, Y_i)\right\|_{\mathcal{G}_{n,\delta}}\right]\right] \\
& \text{using corollary 2.2.8, where } \epsilon_i\text{'s are iid Rademacher r.v's} \\
\lesssim \;& \frac{1}{\sqrt{n}} E_X\left[\int_0^\infty \sqrt{\log(N(\eta, \mathcal{G}_{n,\delta}, L_2(\mathbb{P}_n)))}\, d\eta\right] \\
& \text{from 2.5.3 in VDVW} \\
= \;& \frac{1}{\sqrt{n}} E_X\left[\int_0^{\|F\|_{\mathbb{P}_n,2}} \sqrt{\log(N(\eta, \mathcal{G}_{n,\delta}, L_2(\mathbb{P}_n)))}\, d\eta\right] \\
& \text{because the covering number equals 1 for } \eta > \|F\|_{\mathbb{P}_n,2} \\
\lesssim \;& \frac{1}{\sqrt{n}} E_X\left[\|F\|_{\mathbb{P}_n,2}\int_0^1 \sqrt{\log(N(\eta, \mathcal{G}_{n,\delta}, L_2(\mathbb{P}_n)))}\, d\eta\right] \\
\lesssim \;& \sqrt{\frac{d}{n}} E_X\left[\|F\|_{\mathbb{P}_n,2}\int_0^1 \sqrt{-\log a}\, da\right] \\
& \text{since } \mathcal{G}_{n,\delta} \text{ has VC dimension bounded by a constant multiple of } p \\
\lesssim \;& \sqrt{\frac{d}{n}} \qquad \text{since } E_X\|F\|_{\mathbb{P}_n,2} \le \sqrt{E[F(X,Y)^2]}\text{, latter of which is a constant}
\end{aligned}
\tag{B.130}
$$

Where each implicit constant from line to line does not depend on $n$. The second term can also be bounded: for any $\theta$ with $\|\theta\|_2 = 1$ and $\|\theta - \theta_0\|_2 \le \delta$

$$
\begin{aligned}
& E\left[\left(Y - \frac{\mu}{2}\right)^2 |1(X^T\theta > 0) - 1(X^T\theta_0 > 0)|\right] \\
\le \;& E_X\left[E_\epsilon\left[\left(\frac{3\mu}{2} + |\varepsilon|\right)^2 |1(X^T\theta > 0) - 1(X^T\theta_0 > 0)|\right]\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq E\left[\left(\frac{3\mu}{2}+|\varepsilon|\right)^2\right] E\left[|1(X^T\theta>0)-1(X^T\theta_0>0)|\right] \\
&\lesssim E\left[|1(X^T\theta>0)-1(X^T\theta_0>0)|\right] \\
&\lesssim \|\theta-\theta_0\|_2 \qquad \text{(B.131)}
\end{aligned}
$$

Where the last line is due to Assumption B1, and again all implicit constants are not dependent on $n$. For all small $\delta$, it is possible to find some constant where $\arccos(\theta \cdot \theta_0) \leq C\delta$ for all $\|\theta-\theta_0\|_2 \leq \delta$, and hence the last line $\lesssim \delta$. Combining (B.130) and (B.131) gives

$$
E(\nu_n) \lesssim \sqrt{\sqrt{\frac{d}{n}}+\delta} \qquad \text{(B.132)}
$$

Referring back to (B.126), this yields

$$
EJ(\nu_n, \mathcal{F}_{n,\delta}) \lesssim \sqrt{d} \qquad \text{(B.133)}
$$

Using this upper bound, a final upper bound for the expected norm of $\mathbb{G}_n$ can be obtained. Because the function $x \to x\sqrt{-\log(x)}$ is a strictly increasing function for all sufficiently small $x$, this means that for all sufficiently large $n$ and small $\delta$, we have

$$
\begin{aligned}
E[\|\mathbb{G}_n 1(\Omega_n)\|_{\mathcal{F}_{n,\delta}} 1(\Omega_n)] &\lesssim E[J(\nu_n, \mathcal{F}_{n,\delta})] \\
&\lesssim \sqrt{d} E(\nu_n)\sqrt{-\log E(\nu_n)} \\
&\lesssim \sqrt{d}\left[\left(\sqrt{\frac{d}{n}}+\delta\right)\left(-\log\left(\sqrt{\frac{d}{n}}+\delta\right)\right)\right]^{1/2} \\
&\leq \sqrt{d}\left[\left(\sqrt{\frac{d}{n}}+\delta\right)\log\frac{1}{\delta}\right]^{1/2} \qquad \text{(B.134)}
\end{aligned}
$$

as $\delta \to 0$.

Going back to Theorem B.2, in terms of its notation,

$$\begin{aligned} & E^* \left[ \sup_{\rho_n(\theta,\theta_0)\leq\delta} |(\mathbb{M}_{n,p} - \mathbb{M}_p)(\nu_n) - (\mathbb{M}_{n,p} - \mathbb{M}_p)(\theta_0)| \, 1(\Omega_n) \right] \\ = & \sqrt{\frac{1}{n}} \left\| \|\mathbb{G}\|_{\mathcal{F}_{n,\delta^2}} \right\|_{P,1} \\ \leq & C\sqrt{\frac{d}{n}} \left[ \left( \sqrt{\frac{d}{n}} + \delta^2 \right) \log \frac{1}{\delta} \right]^{1/2} \end{aligned} \tag{B.135}$$

for some constant $C$ independent of $n$. Therefore the $\phi_n$ can be defined as

$$\phi_n(\delta) := C\sqrt{d} \left[ \left( \sqrt{\frac{d}{n}} + \delta^2 \right) \log \frac{1}{\delta} \right]^{1/2} \tag{B.136}$$

and this indeed satisfies $\phi(\delta)/\delta^\alpha$ being a decreasing function for $\alpha \in (1, 2)$. In order for

$$\begin{aligned} \sqrt{n} \geq r_n^2 \phi_n(r_n^{-1}) &= C r_n^2 \sqrt{d} \left[ \left( \frac{d}{n} + \frac{1}{r_n^2} \right) \log r_n \right]^{1/2} \qquad \text{or} \\ \frac{1}{C^2} \frac{n}{d} &\geq r_n^4 \left[ \left( \frac{d}{n} + \frac{1}{r_n^2} \right) \log r_n \right], \end{aligned} \tag{B.137}$$

both of the following are necessary:

$$\begin{aligned} r_n^4 \log r_n &\leq \frac{1}{C^2} \left( \frac{n}{d} \right)^2 \\ r_n^2 \log r_n &\leq \frac{1}{C^2} \cdot \frac{n}{d}. \end{aligned} \tag{B.138}$$

The second line implies the first line, and the second line can be satisfied by having

$$r_n = C^* \sqrt{\frac{n}{d}} \left( \log \frac{n}{d} \right)^{-1/2} \tag{B.139}$$

for some constant $C^*$ not dependent on $n$. Hence, $r_n \rho_n(\theta, \theta_0) = O_p(1)$ means $\|\hat{\theta} - \theta_0\|_2 = O_p\left(\frac{d}{n} \log \frac{d}{n}\right)$.

# BIBLIOGRAPHY

# BIBLIOGRAPHY


Bai, J., and P. Perron (1998), Estimating and testing linear models with multiple structural changes, *Econometrica*, pp. 47–78.

Basseville, M., and I. V. Nikiforov (1993), *Detection of Abrupt Changes: Theory and Application*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Carl, G., G. Kesidis, R. R. Brooks, and S. Rai (2006), Denial-of-service attack-detection techniques, *IEEE Internet computing*, *10*(1), 82–89.

Cho, H., and P. Frylewicz (), Corrections on "multiple change-point detection for high-dimensional time series via sparsified binary segmentation" [1], `https://people.maths.bris.ac.uk/~mahrc/papers/sbs_correction.pdf`, accessed: 2018-12-04.

Cho, H., and P. Fryzlewicz (2015), Multiple-change-point detection for high dimensional time series via sparsified binary segmentation, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *77*(2), 475–507.

Csörgö, M., and L. Horváth (1997), *Limit theorems in change-point analysis*, John Wiley & Sons Inc.

der Vaart, A. W. V., and J. A. Wellner (1996), *Weak Convergence and Empirical Process: With Applications to Statistics*, Springer-Verlag.

Dondelinger, F., S. Mukherjee, and T. A. D. N. Initiative (2016), High-dimensional regression over disease subgroups, *arXiv preprint arXiv:1611.00953*.

Fan, A., R. Song, and W. Lu (2017), Change-plane analysis for subgroup detection and sample size calculation, *Journal of the American Statistical Association*, *112*(518), 769–778.

Fan, J., and Y. Fan (2008), High dimensional classification using features annealed independence rules, *Annals of statistics*, *36*(6), 2605.

Frick, K., A. Munk, and H. Sieling (2014), Multiscale change point inference, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(3), 495–580.

Frisén, M. (2008), *Financial surveillance*, vol. 71, John Wiley & Sons.

Fryzlewicz, P., et al. (2014), Wild binary segmentation for multiple change-point detection, *The Annals of Statistics*, *42*(6), 2243–2281.

Geer, S. A. (2000), *Empirical Processes in M-estimation*, Cambridge university press.

Giraud, C. (2014), *Introduction to high-dimensional statistics*, Chapman and Hall/CRC.

Hansen, B. E. (2000), Sample splitting and threshold estimation, *Econometrica*, *68*(3), 575–603.

Harchaoui, Z., and C. Lévy-Leduc (2010), Multiple change-point estimation with a total variation penalty, *Journal of the American Statistical Association*, *105*(492), 1480–1493.

Huang, T., B. Wu, P. Lizardi, and H. Zhao (2005), Detection of dna copy number alterations using penalized least squares regression, *Bioinformatics*, *21*(20), 3811–3817.

Imai, K., M. Ratkovic, et al. (2013), Estimating treatment effect heterogeneity in randomized program evaluation, *The Annals of Applied Statistics*, *7*(1), 443–470.

Kallitsis, M., S. A. Stoev, S. Bhattacharya, and G. Michailidis (2016), Amon: An open source architecture for online monitoring, statistical analysis, and forensics of multi-gigabit streams, *IEEE Journal on Selected Areas in Communications*, *34*(6), 1834–1848.

KANG, H., C. I. GARCIA, K. CHIN, and A. J. DEARDO (2007), Phd thesis, university of pittsburgh phd thesis, university of pittsburgh, 2004, *ISIJ international*, *47*(3), 486–492.

Khan, N., S. McClean, S. Zhang, and C. Nugent (2016), Optimal parameter exploration for online change-point detection in activity monitoring using genetic algorithms, *Sensors*, *16*(11), 1784.

Killick, R., P. Fearnhead, and I. A. Eckley (2012), Optimal detection of changepoints with a linear computational cost, *Journal of the American Statistical Association*, *107*(500), 1590–1598.

Koepcke, L., G. Ashida, and J. Kretzberg (2016), Single and multiple change point detection in spike trains: Comparison of different cusum methods, *Frontiers in systems neuroscience*, *10*.

Kosorok, M. R. (2007), *Introduction to empirical processes and semiparametric inference*, Springer Science & Business Media.

Lan, Y., M. Banerjee, G. Michailidis, et al. (2009), Change-point estimation under adaptive sampling, *The Annals of Statistics*, *37*(4), 1752–1791.

Li, J., Y. Li, and B. Jin (2018), Multi-threshold change plane model: Estimation theory and applications in subgroup identification, *arXiv preprint arXiv:1808.00647*.

Loader, C. R., et al. (1996), Change point estimation using nonparametric regression, *The Annals of Statistics*, *24*(4), 1667–1678.

Niu, Y. S., and H. Zhang (2012), The screening and ranking algorithm to detect dna copy number variations, *The annals of applied statistics*, *6*(3), 1306.

Niu, Y. S., N. Hao, H. Zhang, et al. (2016), Multiple change-point detection: A selective overview, *Statistical Science*, *31*(4), 611–623.

Pein, F., H. Sieling, and A. Munk (2016), Heterogeneous change point inference, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Qiu, P. (2013), *Introduction to statistical process control*, CRC Press.

Seifert, B., M. Brockmann, J. Engel, and T. Gasser (1994), Fast algorithms for nonparametric curve estimation, *Journal of Computational and Graphical Statistics*, *3*(2), 192–213.

Seo, M. H., and O. Linton (2007), A smoothed least squares estimator for threshold regression models, *Journal of Econometrics*, *141*(2), 704–735.

Shen, Y., R. Lindenbergh, and J. Wang (2016), Change analysis in structural laser scanning point clouds: The baseline method, *Sensors*, *17*(1), 26.

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Venkatraman, E. S. (1992), Consistency results in multiple change-point problems, Ph.D. thesis, to the Department of Statistics.Stanford University.

Wei, S., and M. R. Kosorok (2014), Latent supervised learning for estimating treatment effect heterogeneity.

Yao, Y.-C. (1984), Estimation of a noisy discrete-time step function: Bayes and empirical bayes approaches, *The Annals of Statistics*, pp. 1434–1447.

Zhang, N. R., and D. O. Siegmund (2007), A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data, *Biometrics*, *63*(1), 22–32.