

Change–Point Estimation Under Adaptive Sampling

Yan Lan, Moulinath Banerjee*and George Michailidis†
University of Michigan

August 22, 2007

Abstract

We consider the problem of locating a jump discontinuity (change-point) in a smooth parametric regression model with a bounded covariate. It is assumed that one can sample the covariate at different values and measure the corresponding responses. Budget constraints dictate that a total of n such measurements can be obtained. A multistage adaptive procedure is proposed, where at each stage an estimate of the change point is obtained and new points are sampled from its *appropriately* chosen neighborhood. It is shown that such procedures accelerate the rate of convergence of the least squares estimate of the change-point. Further, the asymptotic distribution of the estimate is derived using empirical processes techniques. The improved efficiency of the procedure is demonstrated using real and synthetic data. This problem is primarily motivated by applications in engineering systems.

Key words and phrases: adaptive sampling, change point estimation, multi–stage procedure, Skorokhod topology, two–stage procedure, zoom-in.

1 Introduction

The problem of estimating the location of a jump discontinuity (change-point) in an otherwise smooth curve has been extensively studied in the nonparametric regression and survival analysis literature; see for example Dempfle and Stute (2002), Gijbels et al. (1999), Gregoire and Hamrouni (2002), Hall and Molchanov (2003), Kosorok and Song (2007), Loader (1996), Mueller (1992), Mueller and Song (1997), Pons (2003), Ritov (1990) and references therein. In the classical setting, measurements on all n covariate–response pairs are available *in advance*, and the main issue is to estimate as accurately as possible the

*Supported in part by NSF Grant DMS-0705288

†Supported in part by NIH grant P41-18627.

location of the change-point. However, there are applications where it is possible to sample the response at any covariate value of the experimenter's choice. The only hard constraint is that the total budget of measurements to be obtained is fixed a priori.

For example, consider the following example from system engineering. There is a stochastic flow of jobs/customers of various types arriving to the system with random service requests. Jobs waiting to be served are placed in queues of infinite capacity. The system's resources are allocated to the various job classes (queues) according to some service policy. This system serves as a canonical queueing model for many applications, including network switches, flexible manufacturing systems, wireless communications, etc. (Hung and Michailidis (2007)). A quantity of great interest to the system's operator is the average delay of the customers, which is a key performance metric of the quality of service offered by the system.

The average delay of the customers in a two class system as a function of its loading, for a resource allocation policy introduced and discussed in Hung and Michailidis (2007), is shown in Figure 1. Specifically, the system was simulated under 134 loading settings and fed by input/service request processes obtained from real network traces and the average delay of 500,000 customers recorded. It can be seen that for loading around 0.8 there is a marked discontinuity in the response, which indicates that under the specified resource allocation policy the service provided to the customers deteriorates. It is of interest to locate the 'threshold' where such a change in the quality of service occurs. It should be pointed out that this threshold would occur at different system loadings for different allocation policies.

A few comments on the setting implied by this example are in order. First, the experimenter can *select* covariate values (in this case the system's loading) and subsequently obtain their corresponding sample responses. Second, the sampled responses are *expensive* to obtain; for example, the average delay is obtained by running a fairly large scale discrete event simulation of the system under consideration, involving half a million customers. For systems, comprised of a large number of customer classes, more computationally intensive simulations that can last days must be undertaken. Third, in many situations there is an *a priori fixed budget* of resources; for this example, it may correspond to CPU time, in other engineering applications to emulation time, while in other scientific contexts to real money.

Given the potentially limited budget of points that can be sampled and lack of a priori knowledge about the location of the change-point the following strategy looks

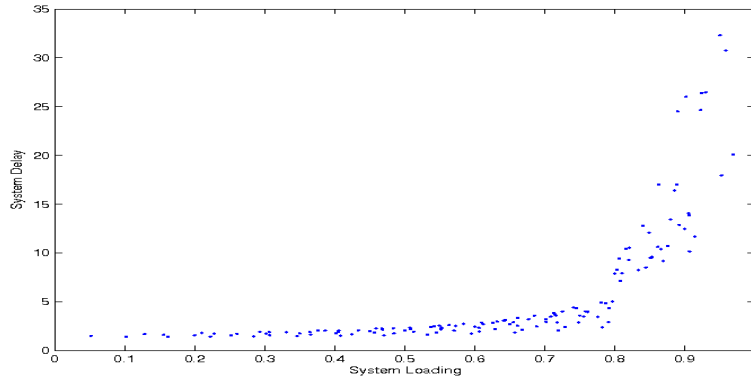


Figure 1: Average delay as a function of system loading for a two-class parallel processing system.

promising. A certain portion of the budget is used to obtain an initial estimate of the change-point based on a least squares criterion. Subsequently, a neighborhood around this initial estimate is specified and the remaining portion of the available points are sampled from it, together with their responses, that yield a new estimate of the change-point. Intuition suggests that if the first stage estimate is fairly accurate, the more intensive sampling in its neighborhood ought to produce a more accurate estimate than the one that would have been obtained by laying out the entire budget of points in a uniform fashion. Obviously, the procedure with its 'zoom-in' characteristics can be extended beyond two stages.

The goal of this paper is to formally introduce such multistage adaptive procedures for change-point estimation and examine their properties. In particular, the following important issues are studied and resolved: (i) the selection of the size of the neighborhoods, (ii) the rate of convergence of the multi-stage least squares estimate, together with its

asymptotic distribution and (iii) allocation of the available budget at each stage.

The proposed procedure should be contrasted with the well studied sequential techniques for change-point detection, since the underlying setting exhibits marked differences. In its simplest form, the sequential change-point detection problem can be formulated as follows: there is a process that generates a sequence of independent observations X_1, X_2, \dots from some distribution F_0 . At some unknown point in time τ , the distribution changes and hence observations $X_\tau, X_{\tau+1}, \dots$ are generated from F_1 . The objective is to raise an alarm as soon as the data generating mechanism switches to new distribution. This problem originally arose in statistical quality control and over the years has found important applications in other fields. Being a canonical problem in sequential analysis, many detection procedures have been proposed in the literature over the years in discrete and continuous time, under various assumptions on the distribution of τ and the data generating mechanism. The literature on this subject is truly enormous; a comprehensive treatment of the problem can be found in the book by Basseville and Nikiforov (1993), while some recent developments and new challenges are discussed in the review paper by Lai (2001). An important difference in our setting is the control that the experimenter exercises over the data generation process and also the absence of physical time, a crucial element in the sequential change-point problem.

The remainder of the paper is organized as follows. In Section 2, the one-stage procedure is briefly reviewed. In Section 3, the proposed procedure based on adaptive sampling is introduced and the main results regarding the rate of convergence of the new estimator of the change point and its asymptotic distribution are established; in addition, various generalizations are discussed. In Section 4, the performance of the proposed estimator in finite samples is studied through an extensive simulation and practical guidelines are discussed for its various tuning parameters. In Section 5, various techniques for constructing confidence intervals for the change-point, based on the two stage procedure, are discussed. Finally, the majority of the technical details are provided in the Appendix.

2 The Classical Problem

In this study, we focus on parametric models for the regression function of the type:

$$Y_i = \mu(X_i) + \epsilon_i, \quad i = 1, 2, \dots, n$$

where

$$\mu(x) = \psi_l(\beta_l, x)1(x \leq d^0) + \psi_u(\beta_u, x)1(x > d^0) \quad (2.1)$$

with $\psi_l(\beta_l, x)$ and $\psi_u(\beta_u, x)$ are both (at least) twice continuously differentiable in β and infinitely differentiable in x and $\psi_l(\beta_l, d^0) \neq \psi_u(\beta_u, d^0)$, so that d^0 is the unique point of discontinuity – a change point – of the regression function.

The ϵ_i 's are assumed to be i.i.d. symmetric mean 0 errors with common (unknown) error variance σ^2 and are independent of the X_i 's which are i.i.d. and are distributed on $[0, 1]$ according to some common density $f_X(\cdot)$. The simplest possible parametric candidate for $\mu(x)$, which we will focus on largely to illustrate the key ideas in the paper, is the simple step function: $\mu(x) = \alpha_0 1(x \leq d^0) + \beta_0 1(x > d^0)$.

Estimating d^0 based on the above data is coined as the ‘‘classical problem’’. A standard way to estimate the parameters (β_l, β_u, d^0) is to solve a least squares problem. We start by introducing some necessary notation. Let \mathbb{P}_n denote the empirical distribution of the data vector $\{X_i, Y_i\}_{i=1}^n$ and P the true distribution of (X_1, Y_1) . For a function f defined on the space $[0, 1] \times \mathbb{R}$ (in which the vector (X_1, Y_1) assumes values) and a measure Q defined on the Borel σ -field on $[0, 1] \times \mathbb{R}$, we denote $\int f dQ$ as Qf . We now turn our attention to the least squares problem.

The objective is to minimize $\mathbb{P}_n [(y - \psi_l(\alpha, x))^2 1(x \leq d) + (y - \psi_u(\beta_u, x))^2 1(x > d)]$ over all (α, β, d) , with $0 \leq d \leq 1$. Let $(\tilde{\beta}_{l,n}, \tilde{\beta}_{u,n}, \tilde{d}_n)$ denote a vector of minimizers. Note that we refer to ‘‘a vector’’ of minimizers, since there will be, in general, multiple tri-vectors that minimize the criterion function. The asymptotic properties of such a vector can be studied by invoking either the methods of Pons (2003) or those (in Chapter 14) of Kosorok (2006) and Kosorok and Song (2006). We do not provide the details, but state the results that are essential to the *multistage learning procedures* that we formulate in the next section. We clarify next the meaning of a minimizer of a right-continuous real-valued function with left limits (say f) defined on an interval I . Specifically, any point $z \in I$ that satisfies $f(z) \wedge f(z-) = \min_{w \in I} f(w)$ is defined to be a minimizer of f . Also, in order to discuss the results for the classical procedure and those for the proposed multistage procedures, we need to define a family of compound Poisson processes that arise in the description of the asymptotic properties of the estimators of the change point.

A family of compound Poisson processes: For a positive constant Λ , let $\nu^+(\cdot)$ be a Poisson process on $[0, \infty)$ with right continuous sample paths, with $\nu^+(s) \sim \text{Poi}(\Lambda s)$ for

$s > 0$. Let $\tilde{\nu}^+(\cdot)$ be another independent Poisson process on $[0, \infty)$ with left-continuous sample paths, with $\tilde{\nu}^+(s) \sim \text{Poi}(\Lambda s)$ and define a (right-continuous) Poisson process on $(-\infty, 0]$ by $\{\nu^-(s) = -\tilde{\nu}^+(-s) : s \in (-\infty, 0]\}$. Let $\{\eta_i^+\}_{i=1}^\infty$ and $\{\eta_i^-\}_{i=1}^\infty$ be two independent sequences of i.i.d. random variables where each η_j (j assumes both positive and negative values) is distributed like η , η being a symmetric random variable with finite variance ρ^2 . Given a positive constant A , define families of random variables $\{V_i^+\}_{i=1}^\infty$ and $\{V_i^-\}_{i=1}^\infty$ where, for each $i \geq 1$, $V_i^+ = A/2 + \eta_i$ and $V_i^- = -A/2 + \eta_{-i}$. Set $V_0^+ = V_0^- \equiv 0$. Next, define compound Poisson processes \mathbb{M}_1 and \mathbb{M}_2 on $(-\infty, \infty)$ as follows: $\mathbb{M}_1(s) = (\sum_{0 \leq i \leq \nu^+(s)} V_i^+) 1(s \geq 0)$ and $\mathbb{M}_2(s) = (\sum_{0 \leq i \leq \nu^-(s)} V_i^-) 1(s \leq 0)$. Finally, define the two-sided compound Poisson process $\mathbb{M}_{A,\eta,\Lambda}(s) = \mathbb{M}_1(s) - \mathbb{M}_2(s)$. It is not difficult to see that \mathbb{M} , almost surely, has a minimizer (in which case it has multiple minimizers, since the sample paths are piecewise constant). Let $d_l(A, \eta, \Lambda)$ denote the smallest minimizer of $\mathbb{M}_{A,\eta,\Lambda}$ and $d_u(A, \eta, \Lambda)$ its largest one, which are, almost surely, well defined. Then, the following relation holds:

$$(d_l(A, \eta, \Lambda), d_u(A, \eta, \Lambda)) \equiv_d \left(d_l \left(\frac{A}{\rho}, \frac{\eta}{\rho}, \Lambda \right), d_u \left(\frac{A}{\rho}, \frac{\eta}{\rho}, \Lambda \right) \right) \quad (2.2)$$

$$\equiv_d \frac{1}{\Lambda} \left(d_l \left(\frac{A}{\rho}, \frac{\eta}{\rho}, 1 \right), d_u \left(\frac{A}{\rho}, \frac{\eta}{\rho}, 1 \right) \right). \quad (2.3)$$

For the "classical problem", the following proposition holds.

Proposition 1: Consider the model described at the beginning of Section 2. Suppose that X has a positive bounded density on $[0, 1]$ and that d^0 is known to lie in the interval $[\epsilon_0, 1 - \epsilon_0]$ for some small $\epsilon_0 > 0$. Let $(\hat{\beta}_{l,n}, \hat{\beta}_{u,n}, \hat{d}_n)$ denote that minimizing tri-vector $(\tilde{\beta}_{l,n}, \tilde{\beta}_{u,n}, \tilde{d}_n)$, for which the third component is minimal. Then, $(\sqrt{n}(\hat{\beta}_{l,n} - \beta_l), \sqrt{n}(\hat{\beta}_{u,n} - \beta_u), n(\hat{d}_n - d^0))$ is $O_p(1)$. Furthermore, the first two components of this vector are asymptotically independent of the third and

$$\begin{aligned} n(\hat{d}_n - d^0) &\rightarrow_d \hat{d}_l(|\mu(d^0+) - \mu(d^0)|, \epsilon_1, f_X(d^0)) \\ &\equiv_d \frac{1}{f_X(d^0)} \hat{d}_l \left(\frac{|\mu(d^0+) - \mu(d^0)|}{\sigma}, \frac{\epsilon_1}{\sigma}, 1 \right). \end{aligned}$$

Heteroscedastic errors: The proposition can be generalized readily to cover the case of heteroscedastic errors. A generalization of the classical model to the heteroscedastic case is as follows: We observe n i.i.d. observations from the model $Y = \mu(X) + \sigma(X) \tilde{\epsilon}$ where $\mu(x)$ is as defined in (2.1), $\tilde{\epsilon}$ and X are independent, $\tilde{\epsilon}$ is symmetric about 0 with unit variance and $\sigma^2(x)$ is a variance function (assumed continuous). As in the homoscedastic case, an unweighted least squares procedure is used to estimate the parameters (β_l, β_u, d^0) .

As before, letting $(\hat{\beta}_{l,n}, \hat{\beta}_{u,n}, \hat{d}_n)$ denote that minimizing tri-vector $(\tilde{\beta}_{l,n}, \tilde{\beta}_{u,n}, \tilde{d}_n)$, for which the third component is minimal, we have:

$$\begin{aligned} n(\hat{d}_n - d^0) &\rightarrow_d d_l(|\mu(d^0+) - \mu(d^0)|, \sigma(d^0) \tilde{\epsilon}, f_X(d^0)) \\ &\equiv_d \frac{1}{f_X(d^0)} d_l\left(\frac{|\mu(d^0+) - \mu(d^0)|}{\sigma(d^0)}, \tilde{\epsilon}, 1\right). \end{aligned}$$

3 The Two-stage Procedure

We first describe a two-stage procedure for estimating the (unique) change-point. In what follows, we consider a regression scenario where the response Y_x generated at covariate level x can be written as $Y_x = \mu(x) + \epsilon$, where ϵ is a symmetric error variable with finite variance and μ is the regression function. The errors corresponding to different covariate levels are i.i.d. We first focus on the simple regression function $\mu(x) = \alpha_0 1(x \leq d^0) + \beta_0 1(x > d^0)$ and discuss generalizations to more complex parametric models later. We are allowed to sample n covariate-response pairs at most and are free to sample a response from any covariate level that we like.

- Step 1: At stage one, λn covariate values are sampled uniformly from $[0, 1]$ and responses are obtained. Denote the observed data by $\{X_i, Y_i\}_{i=1}^{n_1}$, $n_1 = \lambda n$ and the corresponding estimated location of the change point by \hat{d}_{n_1} .
- Step 2: Sample the remaining $n_2 = (1 - \lambda)n$ covariate-response pairs $\{U_i, W_i\}_{i=1}^{n_2}$, where:

$$W_i = \mu(U_i) + \epsilon_i, \quad U_i \sim \text{Unif}[\hat{a}_{n_1}, \hat{b}_{n_1}]$$

and $[\hat{a}_{n_1}, \hat{b}_{n_1}] = [\hat{d}_{n_1} - K n_1^{-\gamma}, \hat{d}_{n_1} + K n_1^{-\gamma}]$, $0 < \gamma < 1$ and K is some constant.

Obtain an updated estimate of the change point based on the n_2 covariate-response pairs from stage 2, which is denoted by \hat{d}_{n_2} .

We discuss the basic procedure in some more detail. Let $(\hat{\alpha}_{n_1}, \hat{\beta}_{n_1}, \hat{d}_{n_1})$ denote the parameter estimates obtained from stage one. Let \mathbb{P}_{n_2} denote the empirical measure of the data points $\{U_i, W_i\}_{i=1}^{n_2}$. The updated estimates are computed by minimizing

$$\mathbb{P}_{n_2}[\{(w - \hat{\alpha}_{n_1})^2 I(u \leq d) + (w - \hat{\beta}_{n_1})^2 I(u > d)]$$

which, as is readily seen, is equivalent to minimizing the process

$$\tilde{\mathbb{M}}_{n_2}(d) \equiv \mathbb{P}_{n_2}[\{(w - \hat{\alpha}_{n_1})^2 - (w - \hat{\beta}_{n_1})^2\}(I(u \leq d) - I(u \leq d^0))].$$

The process \tilde{M}_{n_2} is a piecewise constant right continuous function with left limits. We let $\hat{d}_{n_2,l}$ and $\hat{d}_{n_2,u}$ denote its minimal and maximal minimizers, respectively. Our goal is to determine the joint limit distribution of normalized versions of $(\hat{d}_{n_2,l}, \hat{d}_{n_2,u})$. This is described in the theorems that follow.

Theorem 3.1 *Assume that the error variable ϵ in the regression model has a finite moment generating function in a neighborhood of 0. Then, the random vector $n^{1+\gamma}(\hat{d}_{n_2,l} - d^0, \hat{d}_{n_2,u} - d^0)$ is $O_p(1)$.*

Remark: The proof of this theorem is fairly technical and particularly long and thus deferred to the Appendix. However, a few words regarding the intuition behind the *accelerated rate* of convergence are in order. For simplicity, consider sampling procedures where instead of sampling from a uniform distribution on the interval of interest, sampling takes place on a uniform grid on the interval. The interval from which sampling takes place at the second stage has length $2K n_1^{-\gamma}$. Since the n_2 covariate values are equispaced over this interval, the resolution of the resulting grid at which responses are measured is $O(n_1^{-\gamma}/n_2) = O(n^{-(1+\gamma)})$ and this determines the rate of convergence of the two stage estimator (just as the rate of convergence in the classical procedure where n covariates are equispaced over $[0, 1]$ is given by the resolution of the resulting grid in that situation, which is simply (n^{-1})).

We next describe the limit distributions of the normalized estimates considered in Theorem 3.1.

Theorem 3.2 *Set $C(K, \lambda, \gamma) = (2K)^{-1} (\lambda/(1 - \lambda))^\gamma$. The random vector $n_2^{1+\gamma}(\hat{d}_{n_2,l} - d^0, \hat{d}_{n_2,u} - d^0)$ converges in distribution to*

$$(d_l(|\alpha_0 - \beta_0|, \epsilon, C(K, \lambda, \gamma)), d_u(|\alpha_0 - \beta_0|, \epsilon, C(K, \lambda, \gamma))).$$

Remark: The asymptotic distributions of the 'zoom-in' estimators are given by the minimizers of a compound Poisson process. The underlying Poisson process is basically the limiting version of the count process $\{\mathcal{P}_n(s) : s \in \mathbb{R}\}$, where $\mathcal{P}_n(s)$ counts the number of U_i 's in the interval $(d^0, d^0 + s/n_2^{1+\gamma}] \cup (d^0 + s/n_2^{1+\gamma}, d^0]$. It can be readily checked that marginally, $\mathcal{P}_n(s)$, converges in distribution to a Poisson random variable with mean $C(K, \lambda, \gamma)s$, using the Poisson approximation to the Binomial distribution. On the other hand, the size of the jumps of the compound Poisson process is basically determined by $|\alpha_0 - \beta_0|/\sigma$, the signal-to-noise ratio in the model.

General parametric models: These results admit ready extensions to the case where the

function $\mu(x)$ is as defined in (2.1). As in the case of a piecewise constant μ , $n_1 \equiv \lambda n$ points are initially used to obtain least squares estimates of (β_l, β_u, d^0) , which we denote by $(\hat{\beta}_{l,n_1}, \hat{\beta}_{u,n_1}, \hat{d}_{n_1})$. Step 2 of the two-stage procedure is identical and the updated estimate \hat{d}_{n_2} is computed by minimizing the criterion function

$$\mathbb{P}_{n_2}[\{(w - \psi_l(\hat{\beta}_{l,n_1}, u))^2 I(u \leq d) + (w - \psi_u(\hat{\beta}_{u,n_1}, u))^2 I(u > d)]$$

which is equivalent to minimizing

$$\tilde{\mathbb{M}}_{n_2}(d) = \mathbb{P}_{n_2}[\{(w - \psi_l(\hat{\beta}_{l,n_1}, u))^2 - (w - \psi_u(\hat{\beta}_{u,n_1}, u))^2\}(I(u \leq d) - I(u \leq d^0))].$$

Letting $\hat{d}_{n_2,l}$ and $\hat{d}_{n_2,u}$ denote the smallest and largest argmins of $\tilde{\mathbb{M}}_{n_2}$ respectively (as in the piecewise constant function case), we have the following Proposition.

Proposition 2: The random vector $n_2^{1+\gamma}(\hat{d}_{n_2,l} - d^0, \hat{d}_{n_2,u} - d^0)$ converges in distribution to

$$(d_l(|\psi_l(\beta_l, d^0) - \psi_u(\beta_u, d^0)|, \epsilon, C(K, \lambda, \gamma)), d_u(|\psi_l(\beta_l, d^0) - \psi_u(\beta_u, d^0)|, \epsilon, C(K, \lambda, \gamma))).$$

The heteroscedastic case: Similar results continue to hold for a heteroscedastic regression model. We formulate the heteroscedastic setting as follows. At any given covariate level x , the observed response $Y_x = \mu(x) + \sigma(x) \tilde{\epsilon}$ with $\mu(x)$ as defined in (2.1), $\sigma^2(x)$ is a (continuous) variance function and $\tilde{\epsilon}$ is a symmetric error variable with unit variance. The errors corresponding to different covariate values are independent. Using the same two stage procedure as described above, the following proposition obtains.

Proposition 3: We have

$$n_2^{1+\gamma}(\hat{d}_{n_2,l} - d^0, \hat{d}_{n_2,u} - d^0) \rightarrow_d (d_l(|\psi_l(\beta_l, d^0) - \psi_u(\beta_u, d^0)|, \sigma(d^0) \tilde{\epsilon}, C(K, \lambda, \gamma)), d_u(|\psi_l(\beta_l, d^0) - \psi_u(\beta_u, d^0)|, \sigma(d^0) \tilde{\epsilon}, C(K, \lambda, \gamma))).$$

Remark: With choice of a constant variance function, $\sigma^2(x) \equiv \sigma^2$, the heteroscedastic model reduces to the homoscedastic one. We, nevertheless, present results for these two situations separately. We also subsequently derive our results for the homoscedastic case, the derivations extending almost trivially to the heteroscedastic case.

3.1 Some Generalizations

We briefly discuss some generalizations of the two–stage procedure. The first of these considers more general neighborhoods of the initial estimate of the change–point and the second generalizes the two–stage procedure to multiple stages.

More general neighborhoods: Instead of considering a polynomially decaying neighborhood of the initial estimate in Step 2, one could consider a more general neighborhood of the form $[\hat{d}_{n_1} - k_n n^{-1}, \hat{d}_{n_1} + k_n n^{-1}]$ where $k_n = o(n^{-1})$ and $k_n \rightarrow \infty$. In this case $(\hat{d}_{n_2,1} - d^0, \hat{d}_{n_2,u} - d^0)$ is $O_p(k_n/n^2)$, so that, in theory, rates logarithmically close to n^2 can be achieved (set $k_n = \log n$). The procedure discussed at the beginning of Section 3 is a special case of this scenario with $k_n = K \lambda^{-\gamma} n^{1-\gamma}$. For the heteroscedastic model described above,

$$\frac{n^2}{k_n} \left(\hat{d}_{n_2,1} - d^0, \hat{d}_{n_2,u} - d^0 \right) \rightarrow_d \left(d_l (| \psi_l(\beta_l, d^0) - \psi_u(\beta_u, d^0) |, \sigma(d^0) \tilde{\epsilon}, (1 - \lambda)/2), \right. \\ \left. d_u (| \psi_l(\beta_l, d^0) - \psi_u(\beta_u, d^0) |, \sigma(d^0) \tilde{\epsilon}, (1 - \lambda)/2) \right) .$$

Multi–stage procedures: Consider a generalization of the two stage procedure to k stages in the setting of the heteroscedastic model with a general parametric regression function μ . Let $\lambda_1, \lambda_2, \dots, \lambda_k$ be the proportions of points used at each stage (where $\lambda_1 + \lambda_2 + \dots + \lambda_k = 1$) and let $n_i = \lambda_i n$. Also, fix sequences of numbers $0 < \gamma_{(k-1)} < \dots < \gamma_{(1)} < 1$ and K_1, K_2, \dots, K_{k-1} (with $K_i > 0$). Having used n_1 points to construct the initial estimate \hat{d}_{n_1} , in the q th ($2 \leq q \leq k$) stage, define the sampling neighborhood as $[\hat{d}_{n_{q-1}} - K_{q-1} n_{q-1}^{-((q-2)+\gamma_{q-1})}, \hat{d}_{n_{q-1}} + K_{q-1} n_{q-1}^{-((q-2)+\gamma_{q-1})}]$, sample n_q covariate–response pairs $\{w_i, u_i\}_{i=1}^{n_q}$ from this neighborhood: $W_i = \mu(U_i) + \epsilon_i$ and update the estimate of the change–point to \hat{d}_{n_q} . Let $(\hat{d}_{n_k,l}, \hat{d}_{n_k,u})$ denote the smallest and largest estimates at stage k . It can be shown that $n_k^{(k-1)+\gamma_{(k-1)}}((\hat{d}_{n_k,l} - d^0), (\hat{d}_{n_k,u} - d^0))$ is $O_p(1)$ and converges in distribution to (d_l, d_u) , where (d_l, d_u) is the vector of the smallest and the largest argmins of the process $\mathbb{M}(| \psi_l(\beta_l, d^0) - \psi_u(\beta_u, d^0) |, \sigma(d^0) \tilde{\epsilon}, C_k)$, with $C_k = (1/2K_{k-1})(\lambda_{k-1}/\lambda_k)^{((k-2)+\gamma_{(k-1)})}$.

3.2 Proof of Theorem 3.2

For the proof of this theorem (and the proof of Lemma 3.2 in the Appendix) we denote the process $\mathbb{M}_{|\alpha_0 - \beta_0|, \epsilon, C(K, \lambda, \gamma)}$ simply by \mathbb{M} and its smallest and largest minimizers simply

by (d_l, d_u) . Our proof of this theorem will rely on continuous mapping for the argmin functional. For the sake of concreteness, in what follows, we assume that $\alpha_0 < \beta_0$. Under this assumption, with probability increasing to 1 as n (and consequently n_1) goes to infinity, $\hat{\alpha}_{n_1} < \hat{\beta}_{n_1}$ and d^0 belongs to the set $[\hat{d}_{n_1} - K n_1^{-\gamma}, \hat{d}_{n_1} + K n_1^{-\gamma}]$. On this set $(\hat{d}_{n_2,l}, \hat{d}_{n_2,u})$ can be obtained by minimizing (the equivalent) criterion function:

$$\mathbb{P}_{n_2} \left[\left(w - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2} \right) (I(u \leq d) - I(u \leq d^0)) \right]$$

and d^0 is characterized as:

$$d^0 = \operatorname{argmin} \mathbf{P} \left[\left(w - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2} \right) (I(u \leq d) - I(u \leq d^0)) \right]$$

where \mathbf{P} is the distribution of (W, U) . Therefore, in what follows, we take:

$$\tilde{\mathbb{M}}_{n_2}(d) = \mathbb{P}_{n_2} \left[\left(w - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2} \right) (I(u \leq d) - I(u \leq d^0)) \right],$$

and $\hat{d}_{n_2,l}$ and $\hat{d}_{n_2,u}$ to be the smallest and largest argmins of this stochastic process. Set $(\xi_{n,l}, \xi_{n,u}) = n_2^{1+\gamma}(\hat{d}_{n_2,l} - d^0, \hat{d}_{n_2,u} - d^0)$. Then $(\xi_{n,l}, \xi_{n,u})$ is the vector of smallest and largest argmins of the stochastic process:

$$\begin{aligned} \mathbb{M}_{n_2}(s) &= \sum_{i=1}^{n_2} \left[\left(w_i - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2} \right) \left(I \left(u_i \leq d^0 + \frac{s}{n_2^{1+\gamma}} \right) - I(u_i \leq d^0) \right) \right] \\ &= \mathbb{M}_{n_2}^+(s) - \mathbb{M}_{n_2}^-(s) \end{aligned}$$

where

$$\mathbb{M}_{n_2}^+(s) = \mathbb{M}_{n_2}(s) \mathbf{1}(s \geq 0) \quad \text{and} \quad \mathbb{M}_{n_2}^-(s) = -\mathbb{M}_{n_2}(s) \mathbf{1}(s \leq 0).$$

We now introduce some notation that is crucial to the subsequent development. Let \mathcal{S} denote the class of piecewise constant right continuous functions with left limits (from \mathbb{R} to \mathbb{R}) that are continuous at every integer point, assume the value 0 at 0 and possess finitely many jumps in every compact interval $[-C, C]$ where $C > 0$ is an integer. Let \tilde{f} denote the pure jump process (of jump size 1) corresponding to the function f ; i.e. \tilde{f} is the piecewise constant right continuous function with left limits, such that for any $s > 0$, $\tilde{f}(s)$ counts the number of jumps of the function f in the interval $[0, s]$, while for $s < 0$, $\tilde{f}(s)$ counts the number of jumps in the set $(s, 0)$.

For any positive integer $C > 0$, let $\mathcal{D}[-C, C]$ denote the class of all right continuous functions with left limits with domain $[-C, C]$ equipped with the Skorokhod topology and let $\mathcal{D}[-C, C] \times \mathcal{D}[-C, C]$ denote the corresponding product space. Finally, let \mathcal{D}_C^0 denote the (metric) subspace of $\mathcal{D}([-C, C]) \times \mathcal{D}([-C, C])$ that comprises all function pairs of the form $(f|_{[-C, C]}, \tilde{f}|_{[-C, C]})$ for $f \in \mathcal{S}$. We have the following Lemma which is proved in the Appendix.

Lemma 3.1 *Let $\{f_n\}$ and f_0 be functions in \mathcal{S} , such that for every positive integer C , $(f_n|_{[-C, C]}, \tilde{f}_n|_{[-C, C]})$ converges to $(f_0|_{[-C, C]}, \tilde{f}_0|_{[-C, C]})$ in \mathcal{D}_0^C where f_0 satisfies the property that no two flat stretches of f_0 have the same height. Let $l_{n,C}$ and $u_{n,C}$ denote the smallest and the largest minimizers of f_n on $[-C, C]$, and $l_{0,C}$ and $u_{0,C}$ denote the corresponding functionals for f_0 . Then $(l_{n,C}, u_{n,C}) \rightarrow (l_{0,C}, u_{0,C})$.*

Consider the sequence of stochastic processes $\mathbb{M}_{n_2}(s)$ and let $\mathbb{J}_{n_2}(s)$ denote the corresponding jump processes. We have:

$$\begin{aligned} \mathbb{J}_{n_2}(s) &= \text{sign}(s) \sum_{i=1}^{n_2} \left[\left(I \left(U_i \leq d^0 + \frac{s}{n_2^{1+\gamma}} \right) - I(U_i \leq d^0) \right) \right] \\ &= \mathbb{J}_{n_2}^+(s) + \mathbb{J}_{n_2}^-(s) \end{aligned}$$

where

$$\mathbb{J}_{n_2}^+(s) = \mathbb{J}_{n_2}(s) 1(s \geq 0) \quad \text{and} \quad \mathbb{J}_{n_2}^-(s) = \mathbb{J}_{n_2}(s) 1(s \leq 0).$$

The jump process corresponding to $\mathbb{M}(s)$ is denoted by $\mathbb{J}(s)$ and is given by $\nu^+(s)1(h \geq 0) + \nu^-(s)1(h \leq 0)$. For each $n, \{\mathbb{M}_{n_2}(s) : s \in \mathbb{R}\}$ lives in \mathcal{S} with probability one. Also, with probability 1, $\{\mathbb{M}(s) : s \in \mathbb{R}\}$ lives in \mathcal{S} . Also, on a set of probability one (which does not depend on C), for every positive integer C , $((\mathbb{M}_{n_2}(s), \mathbb{J}_{n_2}(s)) : s \in [-C, C])$ belongs to \mathcal{D}_0^C and so does $((\mathbb{M}(s), \mathbb{J}(s)) : s \in [-C, C])$. Let $(\xi_{n,C,l}, \xi_{n,C,u})$ denote the smallest and largest argmin of \mathbb{M}_{n_2} restricted to $[-C, C]$ and let $(d_{C,l}, d_{C,u})$ denote the corresponding functionals for \mathbb{M} restricted to $[-C, C]$. We prove in the Appendix:

Lemma 3.2 *For every $C > 0$, $((\mathbb{M}_{n_2}(s), \mathbb{J}_{n_2}(s)) : s \in [-C, C])$ converges in distribution to $((\mathbb{M}(s), \mathbb{J}(s)) : s \in [-C, C])$ in the space \mathcal{D}_C^0 .*

Consider the function h that maps an element (a pair of functions) of \mathcal{D}_0^C to the two dimensional vector given by the smallest argmin and the largest argmin of the first component of the element. Using the fact that almost surely no two flat stretches of \mathbb{M} have the same height, it follows by Lemma 3.1 that the process $((\mathbb{M}(s), \mathbb{J}(s)) : s \in [-C, C])$

belongs, almost surely, to the continuity set of the function h . This, coupled with the distributional convergence established in Lemma 3.2 leads to the conclusion that

$$(\hat{\xi}_{n,C,l}, \hat{\xi}_{n,C,u}) \rightarrow_d (d_{C,l}, d_{C,u}). \quad (3.4)$$

We will show that $(\xi_{n,l}, \xi_{n,u}) \rightarrow (d_l, d_u)$. To this end, we use the following lemma from Prakasa Rao (1969).

Lemma 3.3 *Suppose that $\{W_{n\epsilon}\}$, $\{W_n\}$ and $\{W_\epsilon\}$ are three sets of random vectors such that*

(i) $\lim_{\epsilon \rightarrow 0} \limsup_{n \rightarrow \infty} P[W_{n\epsilon} \neq W_n] = 0$, (ii) $\lim_{\epsilon \rightarrow 0} P[W_\epsilon \neq W] = 0$ and (iii) For every $\epsilon > 0$, $W_{n\epsilon} \rightarrow_d W_\epsilon$ as $n \rightarrow \infty$.

Then $W_n \rightarrow_d W$, as $n \rightarrow \infty$.

Before applying the lemma, we first note the following facts: (a) The sequence of (smallest and largest minimizers) $(\xi_{n,l}, \xi_{n,u})$ is $O_p(1)$, and (b) The minimizers (d_l, d_u) are $O_p(1)$. Now, in the above lemma, set $\epsilon = 1/C$, $W_{n\epsilon} = (\xi_{n,C,l}, \xi_{n,C,u})$, $W_\epsilon = (d_{C,l}, d_{C,u})$, $W_n = (\xi_{n,l}, \xi_{n,u})$ and $W = (d_l, d_u)$. Condition (iii) is established in (3.4). From (a) and (b) it follows that Conditions (i) and (ii) of the lemma are satisfied. We conclude that $(\xi_{n,l}, \xi_{n,u}) \rightarrow_d (d_l, d_u)$. \square

Remark: It is instructive to compare the obtained result on the convergence of the (non-unique) argmin functional to that considered in Ferger (2004). Ferger deals with the convergence of the argmax functional under the Skorokhod topology in Theorems 2 and 3 of his paper. Since the argmax functional is *not continuous* under the Skorokhod topology, an exact result on distributional convergence cannot be achieved. Instead, asymptotic upper and lower bounds are obtained on the distribution function of the argmax in terms of the smallest maximizer and the largest maximizer of the limit process (page 88 of Ferger (2004)). The result we obtain here is, admittedly, in a more specialized set-up than the one considered in his paper, but it is stronger since we are able to show *exact* distributional convergence of argmins. This is achieved at the cost of some extra effort: establishing the joint convergence of the original processes, whose argmins are of interest, and their jump processes, and subsequently invoking continuous mapping. Under this stronger mode of convergence, the argmin functional indeed turns out to be continuous, as Lemma 3.1 shows (the arguments employed are similar in spirit to those in section 14.5.1 of Kosorok (2006)). This result allows us to construct asymptotic confidence intervals that have *exact* coverage

at any given level, as opposed to the conservative intervals proposed in Ferger (2004). That the exact confidence intervals buy us significant precision over the conservative ones is evident from the reported simulation results discussed in Section 5.

4 Strategies for Parameter Allocation in Finite Samples

In this section, we describe strategies for selecting the tuning parameters K, γ and λ used in the procedure. We do this in the setting of the simple regression model $\mu(x) = \alpha_0 1(x \leq d^0) + \beta_0 1(x > d^0)$ and homoscedastic normal errors, obvious analogues holding in more general settings.

Recall that $(\hat{d}_{n_2,l}, \hat{d}_{n_2,u})$ are the minimal and maximal minimizers at Step 2. Set: $\hat{d}_{2,av} = (\hat{d}_{n_2,l} + \hat{d}_{n_2,u})/2$. In what follows we use this as our second stage estimate of the change-point. Using notation from Theorem 3.2 of this paper, we have:

$$n_2^{1+\gamma}(\hat{d}_{2,av} - d^0) \rightarrow_d \frac{d_l + d_u}{2}.$$

It is also not difficult to see that this limit distribution is symmetric about 0.

Henceforth, the notation Argmin will denote the simple average of the minimal and maximal minimizers of a compound Poisson process. The quantity $|\alpha_0 - \beta_0|/\sigma$ will be denoted as SNR (signal-to-noise ratio). The higher the SNR, the more advantageous the estimation of the change point at any given sample size would be. By (2.2) and (2.3), we have:

$$\mathbb{M} \equiv \mathbb{M}_{|\alpha_0 - \beta_0|, \epsilon, C(K, \lambda, \gamma)} \equiv_d \frac{1}{C(K, \lambda, \gamma)} \mathbb{M}_{\text{SNR}, Z, 1},$$

where Z is standard normal random variable. This is a consequence of the fact that $\epsilon/\sigma \sim N(0, 1)$. From Theorem 3.2 and the above display, we have:

$$\begin{aligned} n^{1+\gamma}(\hat{d}_{2,av} - d^0) &\rightarrow_d \frac{1}{(1-\lambda)^{1+\gamma}} 2K \left(\frac{1-\lambda}{\lambda}\right)^\gamma \text{Argmin } \mathbb{M}_{\text{SNR}, Z, 1} \\ &\equiv_d \frac{2K}{(1-\lambda)\lambda^\gamma} \text{Argmin } \mathbb{M}_{\text{SNR}, Z, 1}. \end{aligned}$$

For a fixed K and γ this immediately provides an optimal allocation for λ . We should choose λ so as to maximize $\gamma \log \lambda + \log(1 - \lambda)$ which occurs at $\lambda_{opt} = \gamma/(1 + \gamma)$. It can be seen that the approximate standard deviation of $\hat{d}_{2,av}$ is then given by $n^{-(1+\gamma)}(2K\tau/\Psi(\gamma))$ where $\Psi(\gamma) = \gamma^\gamma/(1 + \gamma)^{1+\gamma}$. This is actually decreasing in γ .

Consider now a one-stage procedure with the covariates sampled from a density f_X , with the estimate of the change-point once again chosen to be the simple average of the minimal and maximal minimizers; call this \hat{d}_{av} . In this case, the standard change-point asymptotics in conjunction with (2.2) and (2.3) give:

$$n(\hat{d}_{av} - d^0) \rightarrow_d \frac{1}{f_X(d^0)} \text{Argmin } \mathbb{M}_{\text{SNR},Z,1}.$$

This immediately provides an expression of the asymptotic efficiency of the two-stage procedure with respect to the one-stage (in terms of ratios of approximate standard deviations) given by:

$$\text{ARE}_{2,1}(n) = \frac{n^\gamma \Psi(\gamma)}{2K f_X(d^0)}.$$

In finite samples, how do we choose our K and γ ? For any γ , note that the interval from which sampling at Stage 2 takes place is of the form $[\hat{d}_{n_1} - K n_1^{-\gamma}, \hat{d}_{n_1} + K n_1^{-\gamma}]$. The requirement that this interval contains d^0 with probability increasing to 1 (with increasing n) translates to choosing $K \equiv K_{(n)}$ in such a way that $K_{(n)} n_1^{-\gamma} \approx C_\xi/n_1$ where C_ξ is the upper ξ 'th quantile of the distribution of $\text{Argmin } \mathbb{M}_{\text{SNR},Z,1}$ (which can be shown to be symmetric about 0). Here ξ is a very small fraction, say, .0005. In other words, we want to “zoom in” but not “zoom in” so much that we systematically start missing d^0 in our sampling interval.

Now, setting $K \equiv K_{(n)} = C_\xi/n_1^{1-\gamma}$, writing $n_1 = n\gamma/(1+\gamma)$ and using the form of the function Ψ , we get an approximate formula for the ARE:

$$\text{ARE}_{2,1}(n) \approx \frac{n}{2C_\xi f_X(d^0)} \frac{\gamma}{(1+\gamma)^2}.$$

The latter is maximized at $\gamma = 1$, which corresponds to an allocation of 50% points at Stage 1 (and the remainder at Stage 2) and gives the approximate ARE as:

$$\text{ARE}_{2,1}(n) \approx \frac{n}{8C_\xi f_X(d^0)}. \quad (4.5)$$

It is not difficult to see that the same approximate formula for the ARE holds for some other measures of dispersion, besides the standard deviation. Let

$$\text{ARE}_{2,1,\text{MAD}}(n) \equiv \frac{E |\hat{d}_{av} - d^0|}{E |\hat{d}_{2,av} - d^0|}$$

and

$$\text{ARE}_{2,1,\text{IQR}}(n) \equiv \frac{\text{IQR}(\hat{d}_{av})}{\text{IQR}(\hat{d}_{2,av})},$$

where both first and second stage estimates are based on samples of size n , and $\text{IQR}(X)$ denotes the inter–quartile range of the distribution of a random variable X . Then, following similar steps to those involved in calculating the ARE based on standard deviations, we conclude that:

$$\text{ARE}_{2,1,\text{MAD}}(n) \approx \text{ARE}_{2,1,\text{IQR}}(n) \approx \frac{n}{8 C_\zeta f_X(d^0)}.$$

The accuracy of the above approximation is confirmed empirically through a simulation study. The setting involves a change–point model given by

$$y_i = 0.5I(x \leq 0.5) + 1.5I(x > d^0) + \epsilon_i, \quad x_i \in (0, 1). \quad (4.6)$$

The variance σ^2 was chosen so that the SNR defined as $(\beta_0 - \alpha_0)/\sigma = 1, 2, 5$ and 8 and the sample size varies in increments of 50 from 50 to 1500. The results based on an interval corresponding to $\alpha = .0025$ are shown in Figure 2. It can be seen that there is great agreement between the theoretical formula for the ARE and the empirical ARE, especially for the IQR at all SNR levels and to a large extent for the MAD measure. On the other hand, the presence of ‘outliers’ amongst the first stage estimates introduces too much variability, which in turn leads to inaccuracies for the proposed formula with the standard deviation as a measure of efficiency.

Remark: The formula for the ARE in (4.5) says that the ‘agnostic’ two stage procedure (‘agnostic’ since the covariates are sampled uniformly at each stage) will eventually, i.e. with increasing n , surpass any one stage procedure, no matter the amount of background information incorporated about the location of the change–point in the one stage process, so long as there is uncertainty about the exact location. One can think of an ‘oracle–type’ one stage procedure where the experimenter samples the covariates from a density that peaks in a neighborhood of d^0 relative to the uniform density (corresponding to high values of $f_X(d^0)$). The faster convergence rate of the two stage procedure relative to this one stage procedure guarantees that with increasing n , the ARE will always go to infinity. Further, expression (4.5) provides an approximation to the minimal sample size required for the two–stage procedure to outperform the ‘classical’ one, a result verified through simulations (not shown here).

Remark: A uniform density has been considered up to this point for sampling second stage covariate–response pairs. We examine next the case of using an arbitrary sampling density $g_U(\cdot)$ supported on the interval $[\hat{d}_{n_1} - Kn_1^{-\gamma}, \hat{d}_{n_1} + Kn_1^{-\gamma}]$ and symmetric around

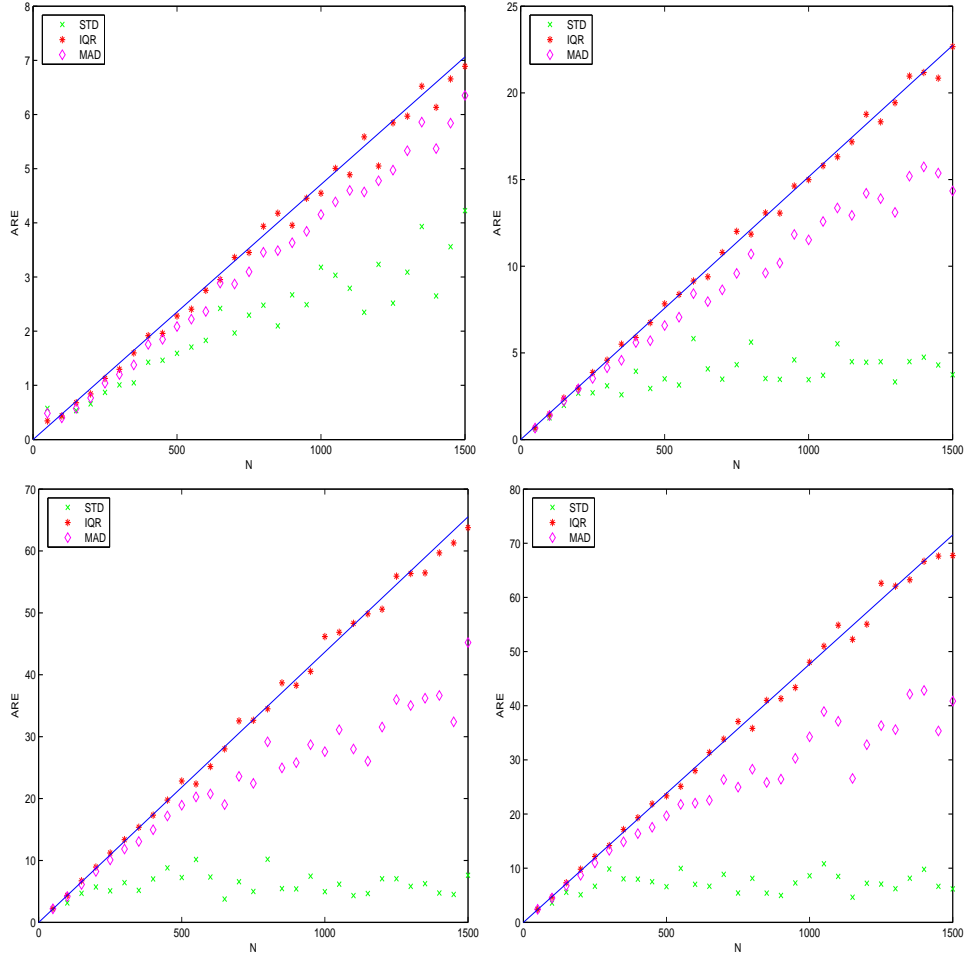


Figure 2: Top panels: ARE for standard deviation, IQR and MAD measures for SNR=1 (left) and 2 (right). Bottom panels: Corresponding ARE for SNR=5 (left) and 8 (right).

\hat{d}_{n-1} . A natural choice for such a density is

$$g_U(d^0) = h\left(\frac{d^0 - \hat{d}_{n_1}}{n_1^{-\gamma} K}\right) \frac{n_1^\gamma}{K},$$

for a density $h(\cdot)$ supported on $[-1, 1]$ and symmetric about 0. Analogous arguments to those used in the proof of Theorem 3.2 establish that the random vector $n_2^{1+\gamma}(\hat{d}_{n_2, l} - d^0, \hat{d}_{n_2, u} - d^0)$ converges in distribution to

$$(d_l(|\alpha_0 - \beta_0|, \epsilon, C(K, \lambda, \gamma, h)), d_u(|\alpha_0 - \beta_0|, \epsilon, C(K, \lambda, \gamma, h))),$$

where

$$C(K, \lambda, \gamma, h) = \left(\frac{\lambda}{1-\lambda} \right)^\gamma \frac{h(0)}{K}.$$

With the error term normally distributed, as assumed in this Section, we get

$$n^{1+\gamma}(\hat{d}_{2,av} - d^0) \rightarrow_d \frac{K}{h(0)(1-\lambda)\lambda^\gamma} \text{Argmin } \mathbb{M}_{\text{SNR},Z,1},$$

and it can be readily checked that the approximate ARE formula reduces to

$$\text{ARE}_{2,1}(n) \approx \frac{nh(0)}{4C_\zeta f_X(d^0)}.$$

It can be seen that the more "peaked" the sampling density g_U (equivalently h) is the greater the efficiency gains. However, one needs to be careful, since the above formula is obtained through asymptotic considerations. In finite samples, a very peaked density around \hat{d}_{n_1} may not perform well, since bias issues (involving $(d^0 - \hat{d}_{n_1})$) must also be taken into account.

4.1 Uniform Sampling Designs

Simulation results indicate that in the presence of a small budget of available points ($n = 20$ or 50), the efficiency of the two-stage estimator can be improved by employing a uniform (equispaced) design in the first stage. The reason is that such a design reduces the sampling variability of the covariate x , which leads to improved localization of the change point. However, the approximate formula for the ARE discussed above is no longer valid when we use a uniform design at the first stage, since the asymptotics of the one stage estimator are then no longer described by the minimizer of a compound Poisson process. We elaborate on this point below.

Suppose that a uniform design on $[0, 1]$ is used in the first stage. The covariate x is then sampled at the points $x_i = i/(n_1 + 1)$, $i = 1, \dots, n_1$. Consider the case where $d^0 = 1/2$, since this is taken up later on in the simulation study. From the results established thus far and the remark following Theorem 3.2, it is clear that the asymptotic distribution of the 1st stage estimator will be determined by the limiting behavior of the count process $\mathcal{P}_n(s)$, where $\mathcal{P}_n(s)$ is once again the number of covariate values in the interval $(d^0, d^0 + s/n_1] \cup (d^0 + s/n_1, d^0]$. However, in the presence of a uniform sampling design, this becomes a *deterministic* process and is given by

$$\lfloor \frac{1}{2}(n_1 + 1) + (1 + \frac{1}{n})s \rfloor - \lfloor \frac{1}{2}(n_1 + 1)d^0 \rfloor \triangleq W_{n_1}(s).$$

From the last expression, it can easily be seen that the limit of $W_{n_1}(s)$ will be different along even and odd subsequences of n_1 for $d^0 = 1/2$ under consideration. As a side remark, notice that for irrational d^0 the limit may not even exist.

Consider $n_1 \rightarrow \infty$ along an even subsequence, so that $W_{n_1}(s) \rightarrow \lfloor 1/2 + |s| \rfloor$. Then, the form of the limit process whose minimizer determines the asymptotic distribution of the least squares estimator is given by $\mathbb{O}(s) = \mathbb{O}_1(s) - \mathbb{O}_2(s)$, with $\mathbb{O}_1(s) = (\sum_{0 \leq i \leq \lfloor 1/2 + |s| \rfloor} V_i^+) 1(s \geq 0)$ and $\mathbb{O}_2(s) = (\sum_{0 \leq i \leq \lfloor 1/2 + |s| \rfloor} V_i^-) 1(s \leq 0)$, with the V_i^\pm defined as before. The main difference is that the number of events up to time s are deterministic and occur at regular intervals, and hence exhibit less variability than a Poisson process with the same number of events up to time s . Therefore, the asymptotics of \hat{d}_{n_1} in this special case are described by the minimizer of a compound process, driven by a deterministic one.

5 Confidence Intervals for the Change-Point

We compare next the performance of exact confidence intervals based on the result established in Theorem 3.2, to those proposed in Ferger (2004). Moreover, confidence intervals for finite samples will be constructed following the discussion in Section 4.

For all these comparisons, simulations were run for a stump model with $\alpha_0 = 0.5$, $\beta_0 = 1.5$, $d^0 = 0.5$ and sample sizes $n = 50, 100, 200, 500, 1000$ with $N = 2000$ replicates for each n . Confidence intervals for d^0 based on the minimal minimizer $\hat{d}_{n_2, l}$, the maximal minimizer $\hat{d}_{n_2, u}$, and the average minimizer $\hat{d}_{n_2, av} = (\hat{d}_{n_2, l} + \hat{d}_{n_2, u})/2$ were constructed. Two values of $\gamma = 1/2$ and $2/3$ and two values of $K = 1$ and 2 were used together with the optimal allocation $\lambda \equiv \gamma/(1 + \gamma)$ as discussed in Section 4. The confidence level was set at $1 - q = .95$ and the percentage of replicates for which the true change-point was included in the corresponding intervals, as well as the average length of each interval, were recorded. In what follows, the symbols d_l and d_u have the same connotations as in the proof of Theorem 3.2.

5.1 Conservative Intervals

Using the results of Ferger (2004), based on any two-stage estimator \hat{d}_{n_2} , we propose an asymptotically conservative confidence interval for d^0 at level $1 - q$:

$$I_n(q) := (\hat{d}_n - b/n_2^{1+\gamma}, \hat{d}_n - a/n_2^{1+\gamma})$$

where $a < b$ are any solutions of the inequality

$$\text{Prob}(d_u < b) - \text{Prob}(d_l \leq a) \geq 1 - q.$$

Based on the smallest, largest and average minimizers at Stage 2, we therefore obtain intervals

$$I_{n_2,l}(q) = (\hat{d}_{n_2,l} - b/n_2^{1+\gamma}, \hat{d}_{n_2,l} - a/n_2^{1+\gamma}),$$

$$I_{n_2,u}(q) = (\hat{d}_{n_2,u} - b/n_2^{1+\gamma}, \hat{d}_{n_2,u} - a/n_2^{1+\gamma}),$$

and

$$I_{n_2,av}(q) = (\hat{d}_{n_2,av} - b/n_2^{1+\gamma}, \hat{d}_{n_2,av} - a/n_2^{1+\gamma})$$

where a is the $q/2^{th}$ quantile of d_l and b is the $(1 - q/2)^{th}$ quantile of d_u . These quantiles do not seem to be analytically determinable but can certainly be simulated to a reasonable degree of approximation.

In the next Table the coverage probabilities together with the length of the confidence intervals are shown for a number of combinations of sample sizes and tuning parameters and with the SNR set equal to 5. It can be seen that the recorded coverage exceeds the nominal

$\gamma = \frac{1}{2}$	n=50		n=100		n=200		n=500		n=1000	
	K=1	K=2	K=1	K=2	K=1	K=2	K=1	K=2	K=1	K=2
$\hat{I}_{n_2,l}$	97.40%	97.55%	97.40%	98.20%	97.65%	98.00%	96.55%	97.05%	97.15%	97.75%
	(.0580)	(.1208)	(.0205)	(.0427)	(.0072)	(.0151)	(.0018)	(.0038)	(.0006)	(.0014)
$\hat{I}_{n_2,u}$	97.05%	98.60%	97.65%	97.05%	97.65%	97.85%	97.40%	97.90%	97.80%	98.00%
	(.0580)	(.1208)	(.0205)	(.0427)	(.0072)	(.0151)	(.0018)	(.0038)	(.0006)	(.0014)
$\hat{I}_{n_2,av}$	99.80%	99.95%	99.80%	99.95%	100%	100%	99.80%	100%	99.70%	99.95%
	(.0580)	(.1208)	(.0205)	(.0427)	(.0072)	(.0151)	(.0018)	(.0038)	(.0006)	(.0014)
$\gamma = \frac{2}{3}$	n=50		n=100		n=200		n=500		n=1000	
$\hat{I}_{n_2,l}$	98.15%	97.70%	97.65%	98.30%	97.90%	97.70%	97.65%	97.30%	97.75%	97.70%
	(.0299)	(.0581)	(.0094)	(.0183)	(.0030)	(.0058)	(.0006)	(.0013)	(.0002)	(.0004)
$\hat{I}_{n_2,u}$	98.00%	98.20%	97.85%	98.55%	97.90%	98.10%	98.30%	97.75%	97.60%	98.50%
	(.0299)	(.0581)	(.0094)	(.0183)	(.0030)	(.0058)	(.0006)	(.0013)	(.0002)	(.0004)
$\hat{I}_{n_2,av}$	99.60%	99.95%	99.60%	99.90%	99.85%	99.95%	99.85%	99.90%	99.85%	99.95%
	(.0299)	(.0581)	(.0094)	(.0183)	(.0030)	(.0058)	(.0006)	(.0013)	(.0002)	(.0004)

Table 1: 95% Conservative Confidence Intervals for a combination of sample sizes and the tuning parameters γ , K and for SNR=5

level of 95% and almost approaching perfect (100%) coverage for the average minimizer.

5.1.1 Exact Confidence Intervals

On the other hand, since Theorem 3.2 provides us with the *exact* asymptotic distributions of the sample minimizers, we can construct asymptotically exact (level $1 - q$ confidence intervals) as follows:

$$\begin{aligned}\tilde{I}_{n_2,l}(q) &= (\hat{d}_{n_2,l} - b_l/n_2^{1+\gamma}, \hat{d}_{n_2,l} - a_l/n_2^{1+\gamma}), \\ \tilde{I}_{n_2,u}(q) &= (\hat{d}_{n_2,u} - b_u/n_2^{1+\gamma}, \hat{d}_{n_2,u} - a_u/n_2^{1+\gamma}), \\ \tilde{I}_{n_2,av}(q) &= (\hat{d}_{n_2,av} - b_{av}/n_2^{1+\gamma}, \hat{d}_{n_2,av} - a_{av}/n_2^{1+\gamma})\end{aligned}$$

where $a_l, b_l, a_u, b_u, a_{av}$ and b_{av} are the exact quantiles (a_l, a_u and a_{av} correspond to $q/2^{th}$ quantiles and b_l, b_u and b_{av} correspond to $(1 - q/2)^{th}$ quantiles) of d_l, d_u and $(d_l + d_u)/2$, respectively.

In the following Table, the coverage probabilities together with the length of the confidence intervals are shown for a number of combinations of sample sizes and tuning parameters and with the SNR set equal to 5. It can be see that the coverage probabilities are

$\gamma = \frac{1}{2}$	n=50		n=100		n=200		n=500		n=1000	
	K=1	K=2	K=1	K=2	K=1	K=2	K=1	K=2	K=1	K=2
$\tilde{I}_{n_2,l}$	95.00%	94.20%	93.80%	95.25%	95.25%	94.10%	93.40%	94.50%	95.05%	94.90%
	(.0283)	(.0599)	(.0100)	(.0212)	(.0035)	(.0075)	(.0009)	(.0019)	(.0003)	(.0007)
$\tilde{I}_{n_2,u}$	94.20%	96.50%	94.80%	94.95%	94.45%	95.85%	94.85%	95.90%	95.50%	95.85%
	(.0294)	(.0602)	(.0104)	(.0213)	(.0037)	(.0075)	(.0009)	(.0019)	(.0003)	(.0007)
$\tilde{I}_{n_2,av}$	94.30%	95.25%	94.35%	94.05%	94.40%	95.45%	93.20%	94.85%	94.55%	95.30%
	(.0236)	(.0487)	(.0083)	(.0172)	(.0029)	(.0061)	(.0007)	(.0015)	(.0003)	(.0005)
$\gamma = \frac{2}{3}$	n=50		n=100		n=200		n=500		n=1000	
$\tilde{I}_{n_2,l}$	95.65%	95.40%	95.05%	96.05%	95.35%	95.45%	95.30%	95.30%	95.05%	95.90%
	(.0148)	(.0277)	(.0047)	(.0087)	(.0015)	(.0027)	(.0003)	(.0006)	(.0001)	(.0002)
$\tilde{I}_{n_2,u}$	95.40%	96.00%	95.65%	96.60%	95.80%	96.15%	96.20%	96.60%	95.15%	96.85%
	(.0149)	(.0302)	(.0047)	(.0095)	(.0015)	(.0030)	(.0003)	(.0006)	(.0001)	(.0002)
$\tilde{I}_{n_2,av}$	95.15%	96.45%	95.20%	96.95%	94.75%	96.10%	95.30%	96.15%	94.05%	96.10%
	(.0120)	(.0253)	(.0038)	(.0080)	(.0012)	(.0025)	(.0003)	(.0005)	(.0001)	(.0002)

Table 2: 95% Exact Confidence Intervals for a combination of sample sizes and tuning parameters γ, K for SNR=5

fairly close to their nominal values, especially for $\gamma = 2/3$. Further, their length is almost half of those obtained according to Ferger's (2004) method. Finally, it should be noted that analogous results were obtained for SNR=2 and 8 (not shown due to space considerations).

5.1.2 Construction of Confidence Intervals in Finite Samples

Confidence intervals in finite samples can also be based on the adaptive parameter allocation strategies discussed in Section 4. We briefly discuss this below, adopting notation from that Section.

With tuning parameters K and γ and the optimal allocation of $\lambda \equiv \gamma/(1 + \gamma)$, a level $1 - q$ confidence interval for d^0 is given by

$$\left[\widehat{d}_{2,av} - n^{-(1+\gamma)} C_{q/2} \frac{2K}{\Psi(\gamma)}, \widehat{d}_{2,av} + n^{-(1+\gamma)} C_{q/2} \frac{2K}{\Psi(\gamma)} \right].$$

With the adaptive choices, $K \equiv K_{(n)} = C_\zeta n_1^{-(1-\gamma)}$, where $q/2 \gg \zeta$ and $n_1 = n\gamma/(1 + \gamma)$, the above confidence interval reduces to:

$$\left[\widehat{d}_{2,av} - \frac{2C_\zeta C_{q/2}}{n^2} \frac{(1 + \gamma)^2}{\gamma}, \widehat{d}_{2,av} + \frac{2C_\zeta C_{q/2}}{n^2} \frac{(1 + \gamma)^2}{\gamma} \right].$$

To minimize the length, we let γ tend to 1, to obtain an approximate level $1 - q$ confidence interval as follows:

$$\left[\widehat{d}_{2,av} - \frac{8C_\zeta C_{q/2}}{n^2}, \widehat{d}_{2,av} + \frac{8C_\zeta C_{q/2}}{n^2} \right].$$

Simulations were run for the above stump model for four different sample sizes: 50, 100, 200, and 500 with 5000 replicates for each sample size and for 3 different values of SNR=2, 5, 8. Confidence intervals as defined above were constructed (with $q = 0.05$). The percentage of intervals containing the true change-point together with their length were recorded and shown in Table 3.

N	SNR=2		SNR=5		SNR=8	
	Coverage	Length	Coverage	Length	Coverage	Length
50	93.24%	0.2780	95.48%	0.0383	95.68%	0.0329
100	94.08%	0.0695	95.24%	0.0096	95.54%	0.0082
200	94.48%	0.0174	94.78%	0.0024	95.16%	0.0021
500	94.82%	0.0028	95.08%	0.00038	94.94%	0.00033

Table 3: 95% Confidence Intervals constructed using the adaptive parameter allocation strategy for different sample sizes and SNR with $\zeta = 0.0005$

We examine next the performance of confidence intervals in finite samples, but where a uniform (equispaced) design is used in the first stage (results shown in Table 4) and in both stages (results shown in Table 5). The setting is identical to that used in Table 3. As the discussion in Section 4.1 indicates, it is not obvious how the tuning parameters C_ζ should be chosen in this case; therefore, the same C_ζ value as the one used in Table 3 was employed. It can be seen that a uniform design used in the 1st stage does not improve performance in terms of coverage or length. However, using a uniform design in both stages and setting C_ζ and C_q to the same values as in Table 3 leads to rather conservative confidence intervals, especially for larger sample sizes and higher values of SNR. Notice that the lengths of the confidence intervals are identical to those in Table 3 due to the choice of the tuning parameters C_ζ and C_q . Nevertheless, experience shows that a uniform design used in the 1st stage gives better mean squared errors in small samples, or when d^0 is closer to the boundary of the covariate's support.

N	SNR=2		SNR=5		SNR=8	
	Coverage	Length	Coverage	Length	Coverage	Length
50	93.72%	0.2780	95.14%	0.0383	95.56%	0.0329
100	93.88%	0.0695	95.12%	0.0096	95.20%	0.0082
200	94.62%	0.0174	95.52%	0.0024	95.52%	0.0021
500	94.72%	0.0028	94.96%	0.00038	95.12%	0.00033

Table 4: 95% Confidence Interval constructed using the adaptive parameter allocation strategy for different sample sizes and SNR with $\zeta = 0.0005$ using a uniform design in the 1st Stage

N	SNR=2		SNR=5		SNR=8	
	Coverage	Length	Coverage	Length	Coverage	Length
50	95.06%	0.2780	100.00%	0.0383	100.00%	0.0329
100	96.66%	0.0695	99.98%	0.0096	100.00%	0.0082
200	96.94%	0.0174	100.00%	0.0024	100.00%	0.0021
500	97.32%	0.0028	99.96%	0.00038	100.00%	0.00033

Table 5: 95% Confidence Interval constructed using the adaptive parameter allocation strategy for different sample sizes and SNR with $\zeta = 0.0005$ using a uniform design in both stages

5.2 Data application

We revisit the motivating application and estimate the change-point using both the "classical" and the developed two-stage procedures. The total budget was set to $n = 70$ and the model fitted to the natural logarithm of the delays comprised two linear segments with a discontinuity. Given that the data (134 system loadings and their corresponding average delays) have been collected in advance, a sampling mechanism close in spirit to selecting covariate values from a uniform distribution was employed for both procedures. Specifically, the necessary number of points was drawn from a uniform distribution in the $[0, 1]$ interval and amongst the available 134 loadings the ones closest to the sampled points were selected, together with their corresponding responses. An analogous strategy was used when a uniform design was employed in the 1st stage of the adaptive procedure. For the two-stage procedure, we set $\lambda = 1/2$ and the remaining tuning parameters to those values provided by the adaptive strategy discussed in Section 4, with $\zeta = .0005$. The results of the "classical" procedure, the two-stage adaptive procedure with sampling from a uniform distribution in both stages and from a uniform design in the 1st stage and the uniform distribution in the 2nd stage are depicted in the left, center and right panels of Figure 3, respectively. The depicted fitted regression models are based on the first stage estimates for the two-stage procedure. Further, the sampled points from the two stages are shown as solid (1st stage) and open (2nd stage) circles. It can be seen that the heavier sampling in the neighborhood of the 1st stage estimate of the change-point improves the estimate given the available evidence from all 134 points shown in Figure 1.

The estimated change-point from the "classical" procedure is $\hat{d}_n = .737$ with a 95% confidence interval $(.682, .793)$. Using a uniform distribution in both stages gave an estimate $\hat{d}_{n_2} = .796$ with a 95% confidence interval $(.781, .811)$. On the other hand, a combination of a uniform design in the 1st stage with that of a uniform distribution in the 2nd stage yielded an estimate $\hat{d}_{n_2} = .802$ with a 95% confidence interval $(.787, .817)$. As shown in this case and validated through other data examples, the use of uniform design in the first stage proves advantageous in practice, especially for small samples or in situations where the discontinuity lies fairly close to the boundary of the design region.

6 Concluding Remarks

In this study, a multistage adaptive procedure for estimating a jump discontinuity in a parametric regression model was introduced and its properties investigated. Specifically, it was established that the the rate of convergence of the least squares estimator of the

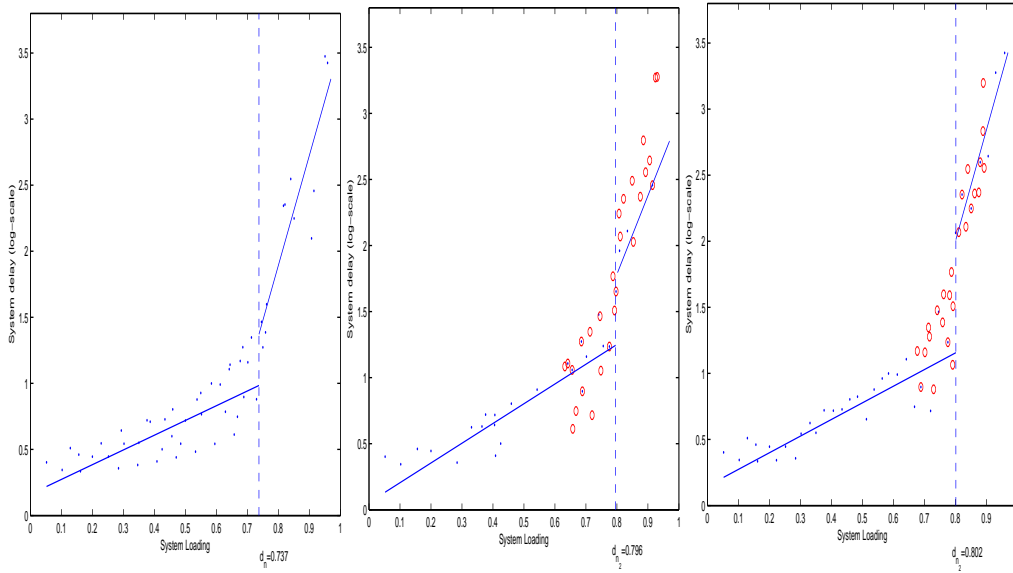


Figure 3: Sampled points (from 1st stage solid circles and from 2nd stage open circles) together with the fitted parametric models and estimated change point, based on a total budget of $n = 70$ points, obtained from the "classical" procedure (left panel), the two-stage adaptive procedure with sampling from a uniform distribution in both stages (center panel) and from a uniform design in the 1st stage and the uniform distribution in the 2nd stage (right panel).

change-point can be accelerated and its asymptotic distribution derived. Several issues pertaining to the tuning of the parameters involved in the procedure were examined and resolved. In practice, it is generally recommended that in the presence of a small budget of points, a uniform design in the first stage proves beneficial. At present the parameters of the parametric model are estimated in the 1st stage of the experiment and held fixed in subsequent stages. One may wonder why the parameters are not re-estimated in the presence of additional samples. The main reason is that the additional samples obtained from a shrinking neighborhood around d^0 are not particularly helpful for estimating global regression parameters. The sole exception is the stump model, where using all n points provide better estimates, especially for small budgets. Finally, it should be noted that a multistage adaptive procedure would also work in the context of a non-parametric model with a jump discontinuity and produce analogous accelerations to the convergence rate of the employed estimator; a comprehensive treatment of this topic is currently under study.

7 Appendix

Proof of Lemma 3.1: Let $f_n(t) = \sum_{i=1}^{\infty} v_{n,i} 1(t \in [a_{n,i}, a_{n,i+1})) + \sum_{i=1}^{\infty} v_{n,-i} 1(t \in [a_{n,-(i+1)}, a_{n,-i}))$ where $0 < a_{n,1} < a_{n,2} < \dots$ and $0 > a_{n,-1} > a_{n,-2} > \dots$, and let $f_0(t) = \sum_{i=1}^{\infty} v_i 1(t \in [a_i, a_{i+1})) + \sum_{i=1}^{\infty} v_{-i} 1(t \in [a_{-(i+1)}, a_{-i}))$. On $[-C, C]$, the limit function f_0 has finitely many jump points, say, m_r to the right of 0 and m_l to the left of 0. The function \tilde{f}_n assumes the value 0 on $[a_{n,-1}, a_{n,1})$ and the value i on $[a_{n,i}, a_{n,i+1})$ and $[a_{n,-(i+1)}, a_{n,-i})$, for all $i \geq 1$. The function \tilde{f}_0 assumes the value 0 on $[a_{-1}, a_1)$ and the value i on $[a_i, a_{i+1})$ and $[a_{-(i+1)}, a_{-i})$, for all $i \geq 1$.

For any $\epsilon > 0$, consider, for $1 \leq i \leq m_r$, the points $a_i - \epsilon, a_i + \epsilon$. Since these are continuity points of f_0 , \tilde{f}_n must converge to \tilde{f}_0 at these finitely many points. Since \tilde{f}_n and \tilde{f}_0 only assume integer values, for all sufficiently large n , $\tilde{f}_n(a_i - \epsilon) = \tilde{f}_0(a_i - \epsilon) = i - 1$ and $\tilde{f}_n(a_i + \epsilon) = \tilde{f}_0(a_i + \epsilon) = i$ for all $1 \leq i \leq m_r$ and $\tilde{f}_0(C) = \tilde{f}_n(C) = m_r$. It follows that the function \tilde{f}_n has exactly m_r jump discontinuities on $[0, C]$ for all sufficiently large n ; furthermore, since \tilde{f}_n jumps between $a_i - \epsilon$ and $a_i + \epsilon$ for all i , $a_{n,i}$, the i 'th largest jump location to the right of 0 satisfies $|a_{n,i} - a_i| \leq \epsilon$ for $1 \leq i \leq m_r$ for n large enough. A similar phenomenon happens to the left of 0, with the number of jumps of \tilde{f}_n in $[-C, 0]$ being exactly m_l for all sufficiently large n , and the jump locations $\{a_{n,-i}\}_{i=1}^{m_l}$ converging to the jump locations $\{a_{-i}\}_{i=1}^{m_l}$ of \tilde{f}_0 on $[-C, 0]$.

Let $[a_j, a_{(j+1)})$ (with $j > 0$) denote the unique stretch on which the restriction of \tilde{f}_0 to $[-C, C]$ is minimized (that the minimizing stretch is unique is guaranteed by our assumptions). The value on this stretch is v_j . Now, consider the points $\{(a_m + a_{m+1})/2 : -m_l \leq m \leq m_r\} \cup \{(-C + a_{-m_l})/2, (a_{m_r} + C)/2\}$. These are continuity points of f_0 (and \tilde{f}_0) and by what has been shown in the previous paragraph, for all sufficiently large n , these are continuity points of \tilde{f}_n with $(a_m + a_{m+1})/2$ lying in the stretch with $a_{n,m}$ and $a_{n,m+1}$ as extremities. Since f_n converges to f_0 in the Skorokhod metric on $[-C, C]$ it follows that $f_n((a_m + a_{m+1})/2) \rightarrow f_0((a_m + a_{m+1})/2)$, for all m . Also $f_n((-C + a_{-m_l})/2)$ and $f_n((a_{m_r} + C)/2)$ converge to $f_0((-C + a_{-m_l})/2)$ and $f_0((a_{m_r} + C)/2)$ respectively. Since v_j is the smallest value of f on $[-C, C]$ and is separated from the remaining possible levels of f_0 on $[-C, C]$ and $v_{n,j}$, the constant value of f_n on the stretch $[a_{n,j}, a_{n,(j+1)})$, converges to v_j , while the constant value of f_n on any other stretch converges to the constant value of f_0 on the corresponding stretch for the limit function, it follows that for all sufficiently large n , v_j^n is separated from the other possible values of f_n and is the smallest value of f_n on $[-C, C]$. It follows that $l_{n,C} = a_{n,j}$ and

$u_{n,C} = a_{n,j+1}$ and these converge to a_j and a_{j+1} which are simply $l_{0,C}$ and $l_{u,C}$. A similar argument works if the stretch on which the minimum of f_0 on $[-C, C]$ is attained lies to the left of 0. \square

Proof of Lemma 3.2: We first note that \mathcal{D}_C^0 (which we view as a metric subspace of $D[-C, C] \times D[-C, C]$) is a measurable subset of $D[-C, C] \times D[-C, C]$. To establish convergence in distribution in the space \mathcal{D}_C^0 , it therefore suffices to establish convergence in distribution in the larger space $D[-C, C] \times D[-C, C]$ (see the discussion in Example 3.1 of Billingsley (1999)). This can be achieved by (a) Establishing finite dimensional convergence: showing that $\{\mathbb{M}_{n_2}(h_i), \mathbb{J}_{n_2}(h_i)\}_{i=1}^l \rightarrow \{\mathbb{M}(h_i), \mathbb{J}(h_i)\}_{i=1}^l$ for all h_1, h_2, \dots, h_l in $[-C, C]$. (b) Verifying tightness of the processes $(\mathbb{M}_{n_2}(h), \mathbb{J}_{n_2}(h))$ under the product topology. But this boils down to verifying marginal tightness.

Let

$$L^+(t) = E[e^{it(w - \frac{\alpha_0 + \beta_0}{2})} \mid U = d^{0+}] \equiv \lim_{d \rightarrow d^{0+}} E[e^{it(w - \frac{\alpha_0 + \beta_0}{2})} \mid U = d]$$

and $L^-(t) = E[e^{it(w - \frac{\alpha_0 + \beta_0}{2})} \mid U = d^0]$. It is not difficult to see that L^+ is the characteristic function of the V_i^+ 's while L^- is the characteristic function of the V_i^- 's. In order to establish finite-dimensional convergence, we first show that for a fixed s , $\mathbb{M}_{n_2}(s)$ converges in distribution to $\mathbb{M}(s)$. We do this via characteristic functions. Consider $\phi_s(t)$, the characteristic function of $\mathbb{M}(s)$ (with $s > 0$). We have:

$$\begin{aligned} E[e^{it\mathbb{M}_1(s)}] &= E[e^{it\sum_{0 \leq k \leq \nu^+(s)} V_k^+}] \\ &= \sum_{l=0}^{\infty} E[e^{it(V_1^+ \dots + V_l^+)}] \frac{e^{-\frac{s}{2K}(\frac{\lambda}{1-\lambda})^\gamma} \left(\frac{s}{2K}(\frac{\lambda}{1-\lambda})^\gamma\right)^l}{l!} \\ &= \sum_{l=0}^{\infty} \frac{(L^+(t))^l \left(\frac{s}{2K}(\frac{\lambda}{1-\lambda})^\gamma\right)^l}{l!} e^{-\frac{s}{2K}(\frac{\lambda}{1-\lambda})^\gamma} \\ &= e^{-\frac{s}{2K}(\frac{\lambda}{1-\lambda})^\gamma} e^{L(t)\frac{s}{2K}(\frac{\lambda}{1-\lambda})^\gamma} \\ &= e^{-\frac{s}{2K}(\frac{\lambda}{1-\lambda})^\gamma(1-L(t))} \end{aligned}$$

We show that $Q_{n_2,s}(t) \equiv E[e^{it\mathbb{M}_{n_2}(s)}]$ converges to $\phi_s(t)$. Let $\xi_{n_1} = n_1(\hat{d}_{n_1} - d^0)$, $\eta_{n_1,1} = \sqrt{n_1}(\hat{\alpha}_{n_1} - \alpha_0)$, $\eta_{n_1,2} = \sqrt{n_1}(\hat{\beta}_{n_1} - \beta_0)$. We have:

$$Q_{n_2,s}(t) = \int Q_{n_2,s}^*(t, \eta_1, \eta_2, \xi) dZ_{n_1}(\eta_1, \eta_2, \xi),$$

where Z_{n_1} is the joint distribution of $(\eta_{n_1,1}, \eta_{n_1,2}, \xi_{n_1})$ and

$$Q_{n_2,s}^*(t, \eta_1, \eta_2, \xi) = E[e^{it\mathbb{M}_{n_2}^+(s)} \mid \eta_{n_1,1} = \eta_1, \eta_{n_1,2} = \eta_2, \xi_{n_1} = \xi]$$

$$= E \left[e^{it \left(W_1 - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2} \right)} \left(I \left(U_1 \leq d^0 + \frac{s}{n_2^{1+\gamma}} \right) - I(U_1 \leq d^0) \right) \Big|_{\eta_{n_1,1} = \eta_1, \eta_{n_1,2} = \eta_2, \xi_{n_1} = \xi} \right]^{n_2}.$$

Let $\epsilon > 0$ be pre-assigned. By Proposition 1, we can find $L > 0$ such that for all sufficiently large n , $Z_{n_1}([-L, L]^3) \geq 1 - \epsilon/3$. Using the fact that characteristic functions are bounded by 1, it follows immediately that for all $n \geq N_0$ (for some N_0),

$$\begin{aligned} |Q_{n_2,s}(t) - \phi_s(t)| &\leq \int_{[-L,L]^3} |Q_{n_2,s}^*(t, \eta_1, \eta_2, \xi) - \phi_s(t)| dZ_{n_1}(\eta_1, \eta_2, \xi) + 2\epsilon/3 \\ &\leq \sup_{(\eta_1, \eta_2, \xi) \in [-L,L]^3} |Q_{n_2,s}^*(t, \eta_1, \eta_2, \xi) - \phi_s(t)| + 2\epsilon/3. \end{aligned}$$

For this fixed L , we now show that for all sufficiently large n

$$D_n \equiv \sup_{(\eta_1, \eta_2, \xi) \in [-L,L]^3} |Q_{n_2,s}^*(t, \eta_1, \eta_2, \xi) - \phi_s(t)| \leq \epsilon/3,$$

whence it follows that eventually $|Q_{n_2,s}(t) - \phi_s(t)| \leq \epsilon$. To show the uniform convergence of $Q_{n_2,s}^*(t, \eta_1, \eta_2, \xi)$ to $\phi_s(t)$ over the compact rectangle $[-L, L]^3$ we proceed as follows.

For given L and C , it is the case that for all sufficiently large n , for any $\xi \in [-L, L]$ and any $0 < s < C$,

$$d^0 + \xi/n_1 - K n_1^{-\gamma} < d^0 < d^0 + s/n_2^{1+\gamma} < d^0 + \xi/n_1 + K n_1^{-\gamma}.$$

Let $P_{n_2,d}(s) \equiv \Pr(d^0 \leq U_1 \leq d^0 + s/n_2^{1+\gamma} \mid \hat{d}_{n_1} = d)$. Consider the conditional characteristic function $Q_{n_2,s}^*(t, \eta_1, \eta_2, \xi)$, for $(\eta_1, \eta_2, \xi) \in [-L, L]^3$. It follows from the above display that for all sufficiently large n (depending only on L and C),

$$\begin{aligned} &Q_{n_2}^*(t, \eta_1, \eta_2, \xi) \\ &= \left[1 - P_{n_2, d^0 + \frac{\xi}{n_1}}(s) \right] + \int_{d^0}^{d^0 + \frac{s}{n_2^{1+\gamma}}} E \left[e^{it \left(W_1 - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2} \right)} \Big|_{U_1 = u} \right] p_{U_1}(u) du \Big]^{n_2} \\ &\quad \left(\text{where } \hat{\alpha}_{n_1} = \alpha_0 + \frac{\eta_1}{\sqrt{n_1}}, \quad \hat{\beta}_{n_1} = \beta_0 + \frac{\eta_2}{\sqrt{n_1}} \right) \\ &= \left[1 - P_{n_2, d^0 + \frac{\xi}{n_1}}(s) \right] + \frac{n_1^\gamma}{2K} \int_{d^0}^{d^0 + \frac{s}{n_2^{1+\gamma}}} E \left[e^{it \left(W_1 - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2} \right)} \Big|_{U_1 = u} \right] du \Big]^{n_2} \\ &= \left[1 - \frac{1}{n_2} \frac{s}{2K} \left(\frac{\lambda}{1-\lambda} \right)^\gamma + \frac{n_1^\gamma}{2K} \int_{d^0}^{d^0 + \frac{s}{n_2^{1+\gamma}}} E \left[e^{it \left(W_1 - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2} \right)} \Big|_{U_1 = u} \right] du \right]^{n_2} \\ &= \left[1 - \frac{1}{n_2} \frac{s}{2K} \left(\frac{\lambda}{1-\lambda} \right)^\gamma + \frac{n_1^\gamma}{2K} \right] \end{aligned}$$

$$\begin{aligned}
& \times \int_0^s E \left[\exp \left\{ it \left(W_1 - \frac{\alpha_0 + \frac{\eta_1}{\sqrt{n_1}} + \beta_0 + \frac{\eta_2}{\sqrt{n_1}}}{2} \right) \right\} \left| U_1 = d^0 + \frac{v}{n_2^{1+\gamma}} \right] \frac{1}{n_2^{1+\gamma}} dv \right]^{n_2} \\
& = \left[1 - \frac{1}{n_2} \frac{s}{2K} \left(\frac{\lambda}{1-\lambda} \right)^\gamma \right. \\
& \quad \left. + \frac{1}{2Kn_2} \left(\frac{\lambda}{1-\lambda} \right)^\gamma e^{-\frac{it(\eta_1+\eta_2)}{2\sqrt{n_1}}} \int_0^s E \left[e^{it(W_1 - \frac{\alpha_0+\beta_0}{2})} \left| U = d^0 + \frac{v}{n_2^{1+\gamma}} \right] dv \right]^{n_2} \right. \\
& \quad \left. + \frac{1}{2Kn_2} \left(\frac{\lambda}{1-\lambda} \right)^\gamma \int_0^s E \left[e^{it(W_1 - \frac{\alpha_0+\beta_0}{2})} \left| U = d^0 + \frac{v}{n_2^{1+\gamma}} \right] dv \right]^{n_2} \right]^{n_2} \\
& = \left[1 - \frac{1}{n_2} \frac{s}{2K} \left(\frac{\lambda}{1-\lambda} \right)^\gamma (1 - B_{n_1, n_2, \eta_1, \eta_2}(s)) \right]^{n_2}
\end{aligned}$$

where

$$B_{n_1, n_2, \eta_1, \eta_2}(s) = \frac{1}{s} e^{-\frac{it(\eta_1+\eta_2)}{2\sqrt{n_1}}} \int_0^s E \left[e^{it(W_1 - \frac{\alpha_0+\beta_0}{2})} \left| U_1 = d^0 + \frac{v}{n_2^{1+\gamma}} \right] dv \right]$$

and

$$D_n = \sup_{(\eta_1, \eta_2) \in [-L, L]^2} \left| \left[1 - \frac{1}{n_2} \frac{s}{2K} \left(\frac{\lambda}{1-\lambda} \right)^\gamma (1 - B_{n_1, n_2, \eta_1, \eta_2}(s)) \right]^{n_2} - \phi_s(t) \right|.$$

Let:

$$z_n(\eta_1, \eta_2) = -\frac{s}{2K} \left(\frac{\lambda}{1-\lambda} \right)^\gamma (1 - B_{n_1, n_2, \eta_1, \eta_2}(s)).$$

It is easy to see that:

$$\tilde{D}_n \equiv \sup_{(\eta_1, \eta_2) \in [-L, L]^2} |z_n(\eta_1, \eta_2) - z_0| \rightarrow 0,$$

where $z_0 = -(s/2K)(\lambda/(1-\lambda))^\gamma(1 - L^+(t))$. Consider now:

$$D_n = \sup_{(\eta_1, \eta_2) \in [-L, L]^2} \left| \left(1 + \frac{1}{n_2} z_n(\eta_1, \eta_2) \right)^{n_2} - e^{z_0} \right|.$$

This is dominated by $I_n + II_n$ where:

$$I_n = \left| \left(1 + \frac{1}{n_2} z_0 \right)^{n_2} - e^{z_0} \right| \rightarrow 0,$$

and

$$II_n = \sup_{(\eta_1, \eta_2) \in [-L, L]^2} \left| \left(1 + \frac{1}{n_2} z_n(\eta_1, \eta_2) \right)^{n_2} - \left(1 + \frac{1}{n_2} z_0 \right)^{n_2} \right|.$$

Since \tilde{D}_n goes to 0, for all sufficiently large n , $|z_0| \vee (\sup_{(\eta_1, \eta_2) \in [-L, L]^2} |z_n(\eta_1, \eta_2)|)$ is bounded by a constant, say M . Straightforward algebra shows that for all sufficiently large

n ,

$$\begin{aligned} II_n &\leq (\sup_{(\eta_1, \eta_2) \in [-L, L]^2} |z_n(\eta_1, \eta_2) - z_0|) \left(\sum_{j=1}^{n_2} \binom{n}{j} \frac{j M^{j-1}}{n_2^j} \right) \\ &= (\sup_{(\eta_1, \eta_2) \in [-L, L]^2} |z_n(\eta_1, \eta_2) - z_0|) \left(1 + \frac{M}{n_2} \right)^{n_2-1} \rightarrow 0. \end{aligned}$$

Thus $D_n \rightarrow 0$ and the uniform convergence of $Q_{n_2}^*(t, \eta_1, \eta_2, \xi)$ to $\phi_s(t) = e^{z_0}$ on $[-L, L]^3$ is established.

We now establish the weak convergence of the finite dimensional distributions of $(\mathbb{M}_{n_2}, \mathbb{J}_{n_2})$ to those of (\mathbb{M}, \mathbb{J}) . For convenience, we restrict ourselves only to the set $[0, C]$. Let J be a positive integer and consider $0 = s_0 < s_1 < s_2 < \dots < s_J \leq C$. Let c_1, c_2, \dots, c_J and d_1, d_2, \dots, d_J be constants. We show that:

$$\begin{aligned} A_n &\equiv E \left(e^{it \sum_{j \leq J} (c_j (\mathbb{M}_{n_2}^+(s_j) - \mathbb{M}_{n_2}^+(s_{j-1})) + d_j (\mathbb{J}_{n_2}^+(s_j) - \mathbb{J}_{n_2}^+(s_{j-1})))} \right) \\ &\rightarrow A \equiv E \left(e^{it \sum_{j \leq J} (c_j (\mathbb{M}(s_j) - \mathbb{M}(s_{j-1})) + d_j (\mathbb{J}(s_j) - \mathbb{J}(s_{j-1})))} \right) \end{aligned}$$

for any vector of constants $(c_1, c_2, \dots, c_J, d_1, d_2, \dots, d_J)$. By the Cramer–Wold device it follows that:

$$\begin{aligned} &(\{\mathbb{M}_{n_2}(s_i) - \mathbb{M}_{n_2}(s_{i-1})\}_{i=1}^J, \{\mathbb{J}_{n_2}(s_i) - \mathbb{J}_{n_2}(s_{i-1})\}_{i=1}^J) \\ &\rightarrow_d (\{\mathbb{M}(s_i) - \mathbb{M}(s_{i-1})\}_{i=1}^J, \{\mathbb{J}(s_i) - \mathbb{J}(s_{i-1})\}_{i=1}^J), \end{aligned}$$

establishing the claim. As before:

$$A_n = \int K_{n_2}^*(t, \eta_1, \eta_2, \xi) dZ_{n_1}(\eta_1, \eta_2, \xi)$$

where

$$\begin{aligned} &K_{n_2}^*(t, \eta_1, \eta_2, \xi) \\ &= E[e^{it \sum_{j \leq J} (c_j (\mathbb{M}_{n_2}^+(s_j) - \mathbb{M}_{n_2}^+(s_{j-1})) + d_j (\mathbb{J}_{n_2}^+(s_j) - \mathbb{J}_{n_2}^+(s_{j-1})))} | \eta_{n_1,1} = \eta_1, \eta_{n_1,2} = \eta_2, \xi_{n_1} = \xi]. \end{aligned}$$

Proceeding as before, the convergence of A_n to A follows if we establish the uniform convergence of $K_{n_2}^*(t, \eta_1, \eta_2, \xi)$ to A on a compact rectangle of the form $[-L, L]^3$. We have:

$$K_{n_2}^*(t, \eta_1, \eta_2, \xi)$$

$$\begin{aligned}
&= E(e^{it \sum_{j \leq J} (c_j (\mathbb{M}_{n_2}^+(s_j) - \mathbb{M}_{n_2}^+(s_{j-1})) + d_j (\mathbb{J}_{n_2}^+(s_j) - \mathbb{J}_{n_2}^+(s_{j-1})))} \mid \eta_{n_1,1} = \eta_1, \eta_{n_1,2} = \eta_2, \xi_{n_1} = \xi) \\
&= E \left[\exp \left\{ it \sum_{j \leq J} \left(\sum_{i=1}^{n_2} \left(c_j \left(W_i - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2} \right) + d_j \right) \left(I \left(U_i \leq d^0 + \frac{s_j}{n_2^{1+\gamma}} \right) \right. \right. \right. \right. \\
&\quad \left. \left. \left. - I(U_i \leq d^0) \right) - \sum_{i=1}^{n_2} \left(c_j \left(W_i - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2} \right) + d_j \right) \left(I \left(U_i \leq d^0 + \frac{s_{j-1}}{n_2^{1+\gamma}} \right) \right) \right. \right. \\
&\quad \left. \left. - I(U_i \leq d^0) \right) \right\} \mid \eta_{n_1,1} = \eta_1, \eta_{n_1,2} = \eta_2, \xi_{n_1} = \xi \right] \\
&= \int E \left[\exp \left\{ it \sum_{j \leq J} \left(\sum_{i=1}^{n_2} \left(c_j \left(W_i - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2} \right) + d_j \right) \left(I \left(U_i \leq d^0 + \frac{s_j}{n_2^{1+\gamma}} \right) \right) \right. \right. \right. \\
&\quad \left. \left. - I \left(U_i \leq d^0 + \frac{s_{j-1}}{n_2^{1+\gamma}} \right) \right) \right\} \mid \eta_{n_1,1} = \eta_1, \eta_{n_1,2} = \eta_2, \xi_{n_1} = \xi \right] \\
&= \int E \left[\exp \left\{ it \sum_{i=1}^{n_2} \left(\sum_{j \leq J} \left(c_j \left(W_i - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2} \right) + d_j \right) \left(I \left(U_i \leq d^0 + \frac{s_j}{n_2^{1+\gamma}} \right) \right) \right. \right. \right. \\
&\quad \left. \left. - I \left(U_i \leq d^0 + \frac{s_{j-1}}{n_2^{1+\gamma}} \right) \right) \right\} \mid \eta_{n_1,1} = \eta_1, \eta_{n_1,2} = \eta_2, \xi_{n_1} = \xi \right] \\
&= \int E \left[\exp \left\{ it \left(\sum_{j \leq J} \left(c_j \left(W_1 - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2} \right) + d_j \right) \left(I \left(U_1 \leq d^0 + \frac{s_j}{n_2^{1+\gamma}} \right) \right) \right. \right. \right. \\
&\quad \left. \left. - I \left(U_1 \leq d^0 + \frac{s_{j-1}}{n_2^{1+\gamma}} \right) \right) \right\} \mid \eta_{n_1,1} = \eta_1, \eta_{n_1,2} = \eta_2, \xi_{n_1} = \xi \right]^{n_2}.
\end{aligned}$$

As previously, for all sufficiently large n (depending possibly only on C and L), for all $\xi \in [-L, L]$,

$$\begin{aligned}
K_{n_2}^*(t, \eta_1, \eta_2, \xi) &= \left[\int_{[d^0, d^0 + s_J/n_2^{1+\gamma}]^C} p_{U_1}(u) du \right. \\
&\quad \left. + \sum_{j \leq J} \int_{d^0 + \frac{s_{j-1}}{n_2^{1+\gamma}}}^{d^0 + \frac{s_j}{n_2^{1+\gamma}}} E \left[e^{it \left(c_j \left(W_1 - \frac{\alpha_0 + \eta_1 / \sqrt{n_1} + \beta_0 + \eta_2 / \sqrt{n_2}} \right) + d_j \right)} \mid U_1 = u \right] p_{U_1}(u) du \right]^{n_2} \\
&= \left[1 - \frac{s_J - s_0}{n_2^{1+\gamma}} \frac{n_1^\gamma}{2K} \right. \\
&\quad \left. + \sum_{j \leq J} \int_{d^0 + \frac{s_{j-1}}{n_2^{1+\gamma}}}^{d^0 + \frac{s_j}{n_2^{1+\gamma}}} E \left[e^{it \left(c_j \left(W_1 - \frac{\alpha_0 + \eta_1 / \sqrt{n_1} + \beta_0 + \eta_2 / \sqrt{n_2}} \right) + d_j \right)} \mid U_1 = u \right] p_{U_1}(u) du \right]^{n_2}
\end{aligned}$$

$$\begin{aligned}
&= \left[1 + \sum_{j \leq J} \left[-\frac{s_j - s_{j-1}}{2K n_2} \left(\frac{\lambda}{1-\lambda} \right)^\gamma + \frac{n_1^\gamma}{2K} \cdot \frac{1}{n_2^{1+\gamma}} \right. \right. \\
&\quad \times \left. \int_{s_{j-1}}^{s_j} E \left[\exp \left\{ it \left(c_j \left(W_1 - \frac{\alpha_0 + \frac{\eta_1}{\sqrt{n_1}} + \beta_0 + \frac{\eta_2}{\sqrt{n_1}} \right) + d_j \right) \right\} \right. \right. \\
&\quad \left. \left. \left| U_1 = d^0 + \frac{v}{1+n_2^{1+\gamma}} \right] dv \right] \right]^{n_2} \\
&= \left[1 + \sum_{j \leq J} \left[-\frac{(s_j - s_{j-1})}{2K n_2} \left(\frac{\lambda}{1-\lambda} \right)^\gamma \right. \right. \\
&\quad \left. \left. + \frac{1}{2K n_2} \left(\frac{\lambda}{1-\lambda} \right)^\gamma (s_j - s_{j-1}) B_{n_1, n_2, \eta_1, \eta_2}^*(s_j, s_{j-1}) \right] \right]^{n_2} \\
&= \left[1 + \sum_{j \leq J} \left[-\frac{(s_j - s_{j-1})}{2K n_2} \left(\frac{\lambda}{1-\lambda} \right)^\gamma (1 - B_{n_1, n_2, \eta_1, \eta_2}^*(s_j, s_{j-1})) \right] \right]^{n_2}
\end{aligned}$$

where

$$\begin{aligned}
&B_{n_1, n_2, \eta_1, \eta_2}^*(s_j, s_{j-1}) \\
&= e^{-it \frac{\eta_1 + \eta_2}{2\sqrt{n_1}}} \int_{s_{j-1}}^{s_j} \frac{1}{s_j - s_{j-1}} E \left[e^{it(c_j(W_1 - \frac{\alpha_0 + \beta_0}{2}) + d_j)} \left| U_1 = d^0 + \frac{v}{n_2^{1+\gamma}} \right] dv
\end{aligned}$$

and

$$\sup_{(\eta_1, \eta_2, \xi) \in [-L, L]^3} |K_{n_2}^*(t, \eta_1, \eta_2, \xi) - A| = \sup_{(\eta_1, \eta_2) \in [-L, L]^2} |K_{n_2}^*(t, \eta_1, \eta_2, \xi) - A|.$$

Setting

$$z_n(\eta_1, \eta_2) = \left[-\sum_{j \leq J} \frac{(s_j - s_{j-1})}{2K} \left(\frac{\lambda}{1-\lambda} \right)^\gamma (1 - B_{n_1, n_2, \eta_1, \eta_2}^*(s_j, s_{j-1})) \right]$$

we see that:

$$\sup_{(\eta_1, \eta_2) \in [-L, L]^2} |z_n(\eta_1, \eta_2) - z_0| \rightarrow 0,$$

where

$$z_0 = \exp \left[-\frac{1}{2K} \left(\frac{\lambda}{1-\lambda} \right)^\gamma \sum_{j \leq J} (s_j - s_{j-1}) (1 - e^{itd_j} L^+(c_j t)) \right],$$

since for each $1 \leq j \leq J$,

$$B_{n_1, n_2, \eta_1, \eta_2}^*(s_j, s_{j-1}) \longrightarrow e^{itd_j} E \left[e^{itc_j \left(w - \frac{\alpha_0 + \beta_0}{2} \right)} \Big| U = d^{0+} \right] = e^{itd_j} L^+(c_j t)$$

uniformly over $(\eta_1, \eta_2) \in [-L, L]^2$. It follows that

$$\begin{aligned} & \sup_{(\eta_1, \eta_2) \in [-L, L]^2} |K_{n_2}^*(t, \eta_1, \eta_2, \xi) - e^{z_0}| \\ &= \sup_{(\eta_1, \eta_2) \in [-L, L]^2} \left| \left(1 + \frac{1}{n_2} z_n(\eta_1, \eta_2) \right)^{n_2} - e^{z_0} \right| \rightarrow 0. \end{aligned}$$

But $e^{z_0} = A$, as can be verified by direct computation. By the property of independent increments of compound Poisson processes, we have:

$$\begin{aligned} & E \left[e^{it \sum_{j \leq J} (c_j (\mathbb{M}(s_j) - \mathbb{M}(s_{j-1})) + d_j (\mathbb{J}(s_j) - \mathbb{J}(s_{j-1})))} \right] \\ &= \prod_{j \leq J} E \left[e^{it(c_j (\mathbb{M}(s_j) - \mathbb{M}(s_{j-1})) + d_j (\mathbb{J}(s_j) - \mathbb{J}(s_{j-1})))} \right] \\ &= \prod_{j \leq J} \sum_{l=0}^{\infty} \sum_{m=0}^l E \left[e^{it(\sum_{0 \leq k \leq l} (c_j V_k^+ + d_j) - \sum_{0 \leq k \leq m} (c_j V_k^+ + d_j))} \right] \cdot P(\nu^+(s_j) = l, \nu^+(s_{j-1}) = m) \\ &= \prod_{j \leq J} \sum_{l=0}^{\infty} \sum_{m=0}^l E \left[e^{it(c_j V_{m+1}^+ + \dots + c_j V_l^+ + (l-m)d_j)} \right] \\ & \quad \times P(\nu^+(s_j) - \nu^+(s_{j-1}) = l - m) \cdot P(\nu^+(s_{j-1}) = m) \\ &= \prod_{j \leq J} \sum_{l=0}^{\infty} \sum_{m=0}^l [e^{itd_j} L^+(c_j t)]^{l-m} \cdot \frac{e^{-\left(\frac{s_j - s_{j-1}}{2K}\right) \left(\frac{\lambda}{1-\lambda}\right)^\gamma}}{(l-m)!} \left[\left(\frac{s_j - s_{j-1}}{2K}\right) \left(\frac{\lambda}{1-\lambda}\right)^\gamma \right]^{l-m} \\ & \quad \times \frac{e^{-\left(\frac{s_j - 1}{2K}\right) \left(\frac{\lambda}{1-\lambda}\right)^\gamma}}{m!} \left[\left(\frac{s_j - 1}{2K}\right) \left(\frac{\lambda}{1-\lambda}\right)^\gamma \right]^m \quad (\text{set } l = m + l') \\ &= \prod_{j \leq J} \sum_{m=0}^{\infty} \frac{e^{-\left(\frac{s_j - 1}{2K}\right) \left(\frac{\lambda}{1-\lambda}\right)^\gamma}}{m!} \left[\left(\frac{s_j - 1}{2K}\right) \left(\frac{\lambda}{1-\lambda}\right)^\gamma \right]^m \sum_{l'=0}^{\infty} ([e^{itd_j} L(c_j t)]^{l'}) \\ & \quad \times \frac{e^{-\left(\frac{s_j - s_{j-1}}{2K}\right) \left(\frac{\lambda}{1-\lambda}\right)^\gamma}}{l'!} \left[\left(\frac{s_j - s_{j-1}}{2K}\right) \left(\frac{\lambda}{1-\lambda}\right)^\gamma \right]^{l'} \\ &= \prod_{j \leq J} e^{-\left(\frac{s_j - s_{j-1}}{2K}\right) \left(\frac{\lambda}{1-\lambda}\right)^\gamma (1 - e^{itd_j} L^+(c_j t))} \\ &= e^{-\sum_{j \leq J} \frac{s_j - s_{j-1}}{2K} \left(\frac{\lambda}{1-\lambda}\right)^\gamma (1 - e^{itd_j} L^+(c_j t))} \\ &\equiv A \end{aligned}$$

This finishes the proof of finite dimensional convergence. This derivation can be extended readily to allow for s_j 's that can also assume negative values; we avoid this here since the derivation involves no new ideas but becomes somewhat more cumbersome.

We finally show that the process \mathbb{M}_{n_2} restricted to $[-C, C]$ is tight. We know that $(\hat{\alpha}_{n_1}, \hat{\beta}_{n_1}, \hat{d}_{n_1}) \xrightarrow{p} (\alpha_0, \beta_0, d^0)$ and $n_1(\hat{d}_{n_1} - d^0) = O_p(1)$. Let

$$\Omega_n = \left\{ |\hat{\alpha}_{n_1} - \alpha_0| \leq \Delta, |\hat{\beta}_{n_1} - \beta_0| \leq \Delta, \hat{d}_{n_1} - \frac{K}{n_1^\gamma} < d^0 - \frac{C}{n_2^{1+\gamma}} < d^0 + \frac{C}{n_2^{1+\gamma}} < d_{n_1} + \frac{K}{n_1^\gamma} \right\}$$

Clearly, $P(\Omega_n) \rightarrow 1$. The event Ω_n can be written as $(\hat{\alpha}_{n_1}, \hat{\beta}_{n_1}, \hat{d}_{n_1}) \in R_n$, where $H_n(R_n) \rightarrow 1$, H_n being the joint distribution of $(\hat{\alpha}_{n_1}, \hat{\beta}_{n_1}, \hat{d}_{n_1})$. Note that $\mathbb{M}_{n_2} \mathbf{1}(\Omega_n)$ is also a process in $D(\mathbb{R})$. We verify tightness of $\mathbb{M}_{n_2} \mathbf{1}(\Omega_n)$ restricted to $[-C, C]$. To this end, we verify (the analogue of) Condition (13.14) on Page 143 of Billingsley (1999), with $\beta = 1/2$ and $\alpha = 1$. Once again, let $0 \leq s_1 \leq s \leq s_2 \leq C$,

$$\begin{aligned} & E[|\mathbb{M}_{n_2}^+(s) - \mathbb{M}_{n_2}^+(s_1)| \cdot |\mathbb{M}_{n_2}^+(s_2) - \mathbb{M}_{n_2}^+(s)| \mathbf{1}(\Omega_n)] \\ &= \int_{R_n} E[|\mathbb{M}_{n_2}^+(s) - \mathbb{M}_{n_2}^+(s_1)| \cdot |\mathbb{M}_{n_2}^+(s_2) - \mathbb{M}_{n_2}^+(s)| | (\hat{\alpha}_{n_1}, \hat{\beta}_{n_1}, \hat{d}_{n_1}) = (\alpha, \beta, d)] dH_n(\alpha, \beta, d) \end{aligned}$$

where

$$\begin{aligned} & E[|\mathbb{M}_{n_2}^+(s) - \mathbb{M}_{n_2}^+(s_1)| \cdot |\mathbb{M}_{n_2}^+(s_2) - \mathbb{M}_{n_2}^+(s)| | (\hat{\alpha}_{n_1}, \hat{\beta}_{n_1}, \hat{d}_{n_1}) = (\alpha, \beta, d)] \\ &= E_{\alpha, \beta, d} \left[\left| \sum_{i=1}^{n_2} \left(W_i - \frac{\alpha + \beta}{2} \right) \left(I \left(U_i \leq d^0 + \frac{s}{n_2^{1+\gamma}} \right) - I \left(U_i \leq d^0 + \frac{s_1}{n_2^{1+\gamma}} \right) \right) \right| \right. \\ & \quad \left. \times \left| \sum_{i=1}^{n_2} \left(W_i - \frac{\alpha + \beta}{2} \right) \left(I \left(U_i \leq d^0 + \frac{s_2}{n_2^{1+\gamma}} \right) - I \left(U_i \leq d^0 + \frac{s}{n_2^{1+\gamma}} \right) \right) \right| \right] \\ &= E_{\alpha, \beta, d} \left[\left| \sum_{i \neq j} I \left(d^0 + \frac{s_1}{n_2^{1+\gamma}} \leq U_i < d^0 + \frac{s}{n_2^{1+\gamma}} \right) I \left(d^0 + \frac{s}{n_2^{1+\gamma}} \leq U_j < d^0 + \frac{s_2}{n_2^{1+\gamma}} \right) \right. \right. \\ & \quad \left. \left(W_i - \frac{\alpha + \beta}{2} \right) \left(W_j - \frac{\alpha + \beta}{2} \right) \right| \right] \\ &\leq \sum_{i \neq j} E_{\alpha, \beta, d} \left[\left| I \left(d^0 + \frac{s_1}{n_2^{1+\gamma}} \leq U_i < d^0 + \frac{s}{n_2^{1+\gamma}} \right) I \left(d^0 + \frac{s}{n_2^{1+\gamma}} \leq U_j < d^0 + \frac{s_2}{n_2^{1+\gamma}} \right) \right. \right. \\ & \quad \left. \left(W_i - \frac{\alpha + \beta}{2} \right) \left(W_j - \frac{\alpha + \beta}{2} \right) \right| \right] \\ &= \sum_{i \neq j} \sup_{t \in [d^0 + s_1 n_2^{-(1+\gamma)}, d^0 + s n_2^{-(1+\gamma)}]} E_{\alpha, \beta, d} \left[\left| W_i - \frac{\alpha + \beta}{2} \right| | U_i = t \right] \frac{n_1^\gamma}{2K} \frac{s - s_1}{n_2^{1+\gamma}} \\ & \quad \times \sup_{t \in [d^0 + s n_2^{-(1+\gamma)}, d^0 + s_2 n_2^{-(1+\gamma)}]} E_{\alpha, \beta, d} \left[\left| W_j - \frac{\alpha + \beta}{2} \right| | U_j = t \right] \frac{n_1^\gamma}{2K} \frac{s_2 - s}{n_2^{1+\gamma}} \end{aligned}$$

$$\begin{aligned}
&\leq \sum_{i \neq j} K_1 \left(\frac{s-s_1}{2Kn_2} \left(\frac{\lambda}{1-\lambda} \right)^\gamma \right) K_2 \left(\frac{s_2-s}{2Kn_2} \left(\frac{\lambda}{1-\lambda} \right)^\gamma \right) \quad \text{for any } (\alpha, \beta, d) \in R_n \\
&= K_1 K_2 \sum_{i \neq j} \frac{(s-s_1)(s_2-s)}{n_2^2 4k^2} \left(\frac{\lambda}{1-\lambda} \right)^{2\gamma} \\
&\leq c^*(s_2-s_1)^2.
\end{aligned}$$

It follows that

$$E[|\mathbb{M}_{n_2}^+(s) - \mathbb{M}_{n_2}^+(s_1)| \cdot |\mathbb{M}_{n_2}^+(s_2) - \mathbb{M}_{n_2}^+(s)| \mathbf{1}(\Omega_n)] \leq H_n(R_n) c^*(s_2-s_1)^2 \leq c^*(s_2-s_1)^2$$

which establishes tightness of $\mathbb{M}_{n_2} \mathbf{1}(\Omega_n)$.

Then, given any $\epsilon > 0, \forall n > N_1$,

$$\text{Prob}[\omega : \mathbb{M}_{n_2} \mathbf{1}(\Omega_n)(\omega) \in \mathcal{K}] \geq 1 - \epsilon$$

where \mathcal{K} is a compact set. But $\text{Prob}[\omega : \omega \in \Omega_n] \geq 1 - \epsilon$ eventually. Therefore, eventually

$$\text{Prob}[\omega \in \Omega_n \text{ and } \mathbb{M}_{n_2} \mathbf{1}(\Omega_n) \in \mathcal{K}] \geq 1 - 2\epsilon$$

and consequently $\text{Prob}[\mathbb{M}_{n_2} \in \mathcal{K}] \geq 1 - 2\epsilon$. This establishes the tightness of \mathbb{M}_{n_2} in the space of right continuous left limits endowed functions on $[-C, C]$. Similarly, the tightness of \mathbb{J}_{n_2} can be established. This completes the verification of marginal tightness and therefore joint tightness. \square

Before embarking on the proof of Theorem 3.1, we need some auxiliary lemmas. We first state these below.

Lemma 7.1 *Let $\tilde{U}_1, \tilde{U}_2, \dots, \tilde{U}_n$ be i.i.d. Uniform (0,1) random variables. Then, for all $\lambda > 0$ and for all $0 < \alpha < \beta \leq 1$, we have:*

$$\text{Pr} \left(\sup_{\alpha \leq s \leq \beta} \frac{|\sqrt{n}(\mathbb{P}_n - P)(1(\tilde{u} \leq s))|}{s} \geq \lambda \right) \leq (\alpha^{-1} - \beta^{-1})\lambda^{-2},$$

where \mathbb{P}_n denotes the empirical measure of the data and P the distribution of \tilde{U}_1 .

This lemma is due to Ferger (2005).

Lemma 7.2 *Suppose that X_1, X_2, \dots, X_n are i.i.d. random elements assuming values in a space \mathcal{X} . Let \mathcal{F} be a class of functions with domain \mathcal{X} and range in $[0, 1]$ with finite*

VC dimension $V(\mathcal{F})$ and set $V = 2(V(\mathcal{F}) - 1)$. Denoting by \mathbb{P}_n the empirical measure corresponding to the sample and by P the distribution of X_1 , we have:

$$Pr^* (\|\sqrt{n}(\mathbb{P}_n - P)\|_{\mathcal{F}} \geq \lambda) \leq \left(\frac{D\lambda}{\sqrt{V}}\right)^V \exp(-2\lambda^2).$$

This lemma is adapted from Talagrand (1994).

Lemma 7.3 *Let $\tilde{U}_1, \tilde{U}_2, \dots, \tilde{U}_n$ be i.i.d. random variables following the uniform distribution on $(0, 1)$. Let $\tilde{\epsilon}_1, \tilde{\epsilon}_2, \dots, \tilde{\epsilon}_n$ be i.i.d. mean 0 random variables with finite variance σ^2 that are independent of the \tilde{U}_i 's. Let $\beta_n(s) = \sum_{i=1}^n \tilde{\epsilon}_i 1(\tilde{U}_i \leq s)$. Then for any $0 < \alpha < \beta < 1$, we have:*

$$Pr \left(\sup_{\alpha \leq s \leq \beta} \frac{|\beta_n(s)|}{s} \geq \lambda \right) \leq (\alpha^{-1} - \beta^{-1})\lambda^{-2}\sigma^2.$$

The proof of this lemma follows the proof of Theorem 3.1.

Lemma 7.4 (Hajek–Renyi inequality) *Consider independent random variables X_1, X_2, \dots , and define $S_n = \sum_{i=1}^n X_i$. Further assume that $E(X_k) = 0$ and $E(X_k^2) < \infty$ for each k . Let $\{c_k\}$ be a decreasing sequence of positive numbers. Then, for any $\epsilon > 0$ and $n, m > 0$ with $n \leq m$, we have:*

$$P(\max_{n \leq k \leq m} |S_k|c_k \geq \epsilon) \leq \frac{1}{\epsilon^2} \left[c_n^2 \sum_{i=1}^n E(X_k^2) + \sum_{k=n+1}^m c_k^2 E(X_k^2) \right].$$

Proof of Theorem 3.1: For simplicity and ease of exposition, in what follows, we assume that n points are used at the first stage to compute estimates $\hat{\alpha}_n, \hat{\beta}_n, \hat{d}_{n_1}$ of the three parameters of interest. At the second stage n i.i.d. U_i 's are sampled from the uniform distribution on $\tilde{D}_n \equiv [\hat{d}_{n_1} - Kn^{-\gamma}, \hat{d}_{n_1} + Kn^{-\gamma}]$ and the updated estimate of d^0 is computed as

$$\hat{d}_{n_2} = \operatorname{argmin}_{u \in \tilde{D}_n} \frac{1}{n} \sum_{i=1}^n [W_i - \hat{\alpha}_n 1(U_i \leq u) - \hat{\beta}_n 1(U_i > u)]^2 \equiv \operatorname{argmin}_{u \in \tilde{D}_n} SS(u).$$

In the above display $W_i = f(U_i) + \epsilon_i$ where ϵ_i 's are i.i.d. error variables. Working under this more restrictive setting (of equal allocation of points at each stage) *does not compromise* the complexity of the arguments involved. Finally, recall that by our assumption, $E[e^{C|\epsilon_1|}]$ is finite, for some $C > 0$.

Before proceeding further, a word about the definition of argmin in the above

display. The function SS is a right-continuous function endowed with left limits. For this derivation, we take the argmin to be the smallest u in \tilde{D}_n for which $\min(SS(u), SS(u-)) = \inf_{x \in \tilde{D}_n} SS(x)$.

Denote by G_n the distribution of $(\hat{\alpha}_n, \hat{\beta}_n, \hat{d}_{n1})$. Now, given $\epsilon > 0$, find L so large that for all sufficiently large n , say $n \geq N_0$,

$$(\hat{\alpha}_n, \hat{\beta}_n, \hat{d}_{n1}) \in [\alpha_0 - L/\sqrt{n}, \alpha_0 + L/\sqrt{n}] \times [\beta_0 - L/\sqrt{n}, \beta_0 + L/\sqrt{n}] \times [d^0 - L/n, d^0 + L/n]$$

with probability greater than $1 - \epsilon$. Denote the region on the right side of the above display by R_n . Then, for all $n \geq N_0$,

$$\begin{aligned} & \Pr(n^{1+\gamma} | \hat{d}_{n2} - d^0 | > a) \\ & \leq \int_{R_n} \Pr(n^{1+\gamma} | \hat{d}_{n2} - d^0 | > a | \hat{\alpha}_n = \alpha, \hat{\beta}_n = \beta, \hat{d}_{n1} = t) dG_n(\alpha, \beta, t) + \epsilon \end{aligned}$$

which is dominated by

$$\sup_{(\alpha, \beta, t) \in R_n} \Pr_{t, \alpha, \beta}(n^{1+\gamma} | \hat{d}_{n2} - d^0 | > a) + \epsilon.$$

By making a large, we will show that for all sufficiently large n (say $n > N_1 > N_0$), the supremum is bounded by ϵ . This will complete the proof.

First note that since N_0 is chosen to be sufficiently large, whenever $n \geq N_0$ and $t \in [d^0 - L/n, d^0 + L/n]$, it is the case that $t - Kn^{-\gamma} < d^0 - Kn^{-\gamma}/2 < d^0 + Kn^{-\gamma}/2 < t + Kn^{-\gamma}$. It is not difficult to see that

$$\begin{aligned} \hat{d}_{n2} &= \operatorname{argmin}_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} \tilde{\mathbb{P}}_n [(w - (\alpha + \beta)/2)(1(u \leq d) - 1(u \leq d^0))] \\ &\equiv \operatorname{argmin}_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} \tilde{\mathbb{M}}_n(d) \end{aligned}$$

and

$$\begin{aligned} d^0 &= \operatorname{argmin}_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} \tilde{P}_n [(w - (\alpha + \beta)/2)(1(u \leq d) - 1(u \leq d^0))] \\ &\equiv \operatorname{argmin}_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} \tilde{M}_n(d), \end{aligned}$$

where \tilde{P}_n is the distribution of the pair (W_1, U_1) generated at stage two under first stage parameters (α, β, t) and $\tilde{\mathbb{P}}_n$ is the empirical measure corresponding to n i.i.d. observations from \tilde{P}_n . Note that

$$\tilde{M}_n(d) = \{ | \beta_0 - (\alpha + \beta)/2 | | d - d^0 | 1(d \geq d^0) + | \alpha_0 - (\alpha + \beta)/2 | | d - d^0 | 1(d < d^0) \} (n^\gamma / 2K).$$

Now, for $0 < r \leq K/2$, set $a(r) = \min \{ \tilde{M}_n(d) : |d - d^0| \geq r n^{-\gamma} \}$. Then $a(r) = \min(|\beta_0 - (\alpha + \beta)/2|, |\alpha_0 - (\alpha + \beta)/2|) r/2K$ and let $b(r) = (a(r) - \tilde{M}_n(d^0))/3 = a(r)/3$. Now, for all $n \geq N_0$, for α, β in the region under consideration, $b(r)$ is readily seen to be uniformly bounded below by κr for some constant κ depending only on $\alpha_0, \beta_0, K, N_0$. We then have:

$$\sup_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} |\tilde{\mathbb{M}}_n(d) - \tilde{M}_n(d)| \leq b(r) \Rightarrow |\hat{d}_{n_2} - d^0| \leq r n^{-\gamma}. \quad (7.7)$$

To prove this, assume that the inequality on the left side of the above display holds and consider $d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]$ with $|d - d^0| > r n^{-\gamma}$. Then,

$$\tilde{\mathbb{M}}_n(d) \geq \tilde{M}_n(d) - b(r) \geq a(r) - b(r)$$

and

$$\tilde{\mathbb{M}}_n(d^0) \leq \tilde{M}_n(d^0) + b(r)$$

$$\Rightarrow \tilde{\mathbb{M}}_n(d) - \tilde{\mathbb{M}}_n(d^0) \geq a(r) - b(r) - \tilde{M}_n(d^0) - b(r) = b(r) > 0$$

Hence

$$\tilde{\mathbb{M}}_n(d) > \tilde{\mathbb{M}}_n(d^0).$$

Now, since \hat{d}_{n_2} is the smallest $d \in \tilde{D}_n$ for which $\tilde{\mathbb{M}}_n(d) \wedge \tilde{\mathbb{M}}_n(d-) = \inf_{x \in \tilde{D}_n} \tilde{\mathbb{M}}_n(x)$ and $\tilde{\mathbb{M}}_n$ is a (right continuous left limits endowed) piecewise constant function with finitely many flat stretches, it is easy to see that $\tilde{\mathbb{M}}_n(\hat{d}_{n_2}) = \inf_{x \in \tilde{D}_n} \tilde{\mathbb{M}}_n(x)$. Therefore, $\tilde{\mathbb{M}}_n(\hat{d}_{n_2}) \leq \tilde{\mathbb{M}}_n(d^0)$, showing that $|\hat{d}_{n_2} - d^0| \leq r n^{-\gamma}$ in view of the last display above.

Now, consider

$$\Pr_{\alpha, \beta, t}(|\hat{d}_{n_2} - d^0| > r n^{-\gamma})$$

$$\leq \Pr_{\alpha, \beta, t}(r n^{-\gamma} < |\hat{d}_{n_2} - d^0| \leq \delta n^{-\gamma}) + \Pr_{\alpha, \beta, t}(|\hat{d}_{n_2} - d^0| > \delta n^{-\gamma}) \quad (7.8)$$

$$\equiv P_n(\alpha, \beta, t) + Q_n(\alpha, \beta, t), \quad (7.9)$$

where δ (is sufficiently small, say less than $K/3$) does not depend on t, α, β . We deal with $Q_n(\alpha, \beta, t)$ later. We first consider $P_n(\alpha, \beta, t) \equiv \Pr_{\alpha, \beta, t}(r n^{-\gamma} < |\hat{d}_{n_2} - d^0| \leq \delta n^{-\gamma})$. Since,

$$\begin{aligned} \{r n^{-\gamma} < |\hat{d}_{n_2} - d^0| \leq \delta n^{-\gamma}\} \subseteq & \left[\bigcup_{d^0 + r n^{-\gamma} < d \leq d^0 + \delta n^{-\gamma}} \{\tilde{\mathbb{M}}_n(d) \leq \tilde{\mathbb{M}}_n(d^0)\} \right] \\ & \cup \left[\bigcup_{d^0 - \delta n^{-\gamma} \leq d < d^0 - r n^{-\gamma}} \{\tilde{\mathbb{M}}_n(d) \leq \tilde{\mathbb{M}}_n(d^0)\} \right], \end{aligned}$$

we conclude that

$$\begin{aligned} P_n(\alpha, \beta, t) &\leq P_{n,1}(\alpha, \beta, t) + P_{n,2}(\alpha, \beta, t) \\ &\equiv \Pr_{\alpha, \beta, t}(\cup_{d^0 + rn^{-\gamma} < d \leq d^0 + \delta n^{-\gamma}} \{\tilde{M}_n(d^0) - \tilde{M}_n(d) \geq 0\}) \\ &\quad + \Pr_{\alpha, \beta, t}(\cup_{d^0 - \delta n^{-\gamma} \leq d < d^0 - rn^{-\gamma}} \{\tilde{M}_n(d^0) - \tilde{M}_n(d) \geq 0\}). \end{aligned}$$

We first construct an upper bound on $\sup_{(\alpha, \beta, t) \in R_n} P_{n,1}(\alpha, \beta, t)$. For any $d \in (d^0 + rn^{-\gamma}, d^0 + \delta n^{-\gamma}]$ we have:

$$\begin{aligned} \tilde{M}_n(d^0) - \tilde{M}_n(d) &= (\tilde{M}_n(d^0) - \tilde{M}_n(d^0)) - (\tilde{M}_n(d) - \tilde{M}_n(d)) - (\tilde{M}_n(d) - \tilde{M}_n(d^0)) \\ &= -(\tilde{M}_n(d) - \tilde{M}_n(d)) - \left| \beta_0 - \frac{\alpha + \beta}{2} \right| \frac{n^\gamma}{2K} |d - d^0|. \end{aligned}$$

Hence:

$$0 \leq \tilde{M}_n(d^0) - \tilde{M}_n(d) \Rightarrow (2K)^{-1} \left| \beta_0 - \frac{\alpha + \beta}{2} \right| \leq \frac{-(\tilde{M}_n(d) - \tilde{M}_n(d))}{n^\gamma |d - d^0|}.$$

Now, for all $(\alpha, \beta, t) \in R_n$ (with $n \geq N_0$), $\left| \beta_0 - \frac{\alpha + \beta}{2} \right|$ is bounded below by some constant B , whence it follows that:

$$0 \leq \tilde{M}_n(d^0) - \tilde{M}_n(d) \Rightarrow \frac{|\tilde{M}_n(d) - \tilde{M}_n(d)|}{n^\gamma |d - d^0|} \geq \frac{B}{2K}.$$

Thus,

$$\begin{aligned} &\cup_{d^0 - \delta n^{-\gamma} < d \leq d^0 - rn^{-\gamma}} \{\tilde{M}_n(d^0) - \tilde{M}_n(d) \geq 0\} \\ &\subset \left\{ \sup_{d^0 + rn^{-\gamma} < d \leq d^0 + \delta n^{-\gamma}} \frac{|\tilde{M}_n(d) - \tilde{M}_n(d)|}{n^\gamma |d - d^0|} \geq \tilde{B} \right\}, \end{aligned}$$

where $\tilde{B} = B/2K$. We thus have:

$$\begin{aligned} P_{n,1}(\alpha, \beta, t) &\leq \Pr_{\alpha, \beta, t} \left[\sup_{d^0 + rn^{-\gamma} < d \leq d^0 + \delta n^{-\gamma}} \frac{|\tilde{M}_n(d) - \tilde{M}_n(d)|}{n^\gamma |d - d^0|} \geq \tilde{B} \right] \\ &= \Pr_{\alpha, \beta, t} \left(\sup_{d^0 + rn^{-\gamma} < d \leq d^0 + \delta n^{-\gamma}} \frac{|\sqrt{n}(\tilde{\mathbb{P}}_n - \tilde{P}_n) f_{d, \alpha, \beta}(u, w)|}{|d - d^0| n^\gamma} \geq \sqrt{n} \tilde{B} \right), \end{aligned}$$

where

$$f_{d, \alpha, \beta}(u, w) = (w - (\alpha + \beta)/2)(1(u \leq d) - 1(u \leq d^0)).$$

Using the fact that for $d > d^0$, $(W_j - (\alpha + \beta)/2)(1(U_j \leq d) - 1(U_j \leq d^0)) = (\beta_0 - (\alpha + \beta)/2)(1(U_j \leq d) - 1(U_j \leq d^0)) + \epsilon_j(1(U_j \leq d) - 1(U_j \leq d^0))$, this upper bound on $P_{n,1}(\alpha, \beta, t)$ is easily seen to be dominated by $I_n + II_n$ where

$$\begin{aligned} &I_n \\ &= \Pr_{\alpha, \beta, t} (|\beta_0 - (\alpha + \beta)/2| \sup_{r < s \leq \delta} \frac{|\sqrt{n}(\tilde{\mathbb{P}}_n - \tilde{P}_n)(1(u \leq d^0 + sn^{-\gamma}) - 1(u \leq d^0))|}{s} \geq \sqrt{n} \tilde{B}/2) \end{aligned}$$

which in turn is dominated by

$$I'_n = \Pr_{\alpha, \beta, t} \left(\sup_{r < s \leq \delta} \frac{|\sqrt{n}(\tilde{\mathbb{P}}_n - \tilde{P}_n)(1(u \leq d^0 + sn^{-\gamma}) - 1(u \leq d^0))|}{s} \geq \frac{\sqrt{n} \tilde{B}}{2B'} \right)$$

where, for $n \geq N_0$ and $(\alpha, \beta, t) \in R_n$, B' is an upper bound on $|\beta_0 - (\alpha + \beta)/2|$, while

$$II_n = \Pr_t \left(\sup_{r < s \leq \delta} \frac{|(1/\sqrt{n}) \sum_{i=1}^n \epsilon_i (1(U_i \leq d^0 + sn^{-\gamma}) - 1(U_i \leq d^0))|}{s} \geq \sqrt{n} \tilde{B}/2 \right).$$

Since the U_i 's are i.i.d. Uniform on $[t - Kn^{-\gamma}, t + Kn^{-\gamma}]$, it is easy to see that I'_n is simply

$$\Pr \left(\sup_{r < s \leq \delta} \frac{|\sqrt{n}(\mathbb{Q}_n - Q)(1(\tilde{w} \leq s))|}{s} \geq \frac{\sqrt{n} \tilde{B}}{2B'} \right)$$

where $\tilde{W}_1, \tilde{W}_2, \dots, \tilde{W}_n$ are i.i.d. $\text{Unif}[0, 2K]$, \mathbb{Q}_n is the empirical measure of the \tilde{W}_i s and Q is the distribution of \tilde{W}_1 . In terms of $\tilde{U}_1, \tilde{U}_2, \dots, \tilde{U}_n$, which are i.i.d. $\text{Unif}[0, 1]$, this expression is simply:

$$\Pr \left(\sup_{r/2K < s \leq \delta/2K} \frac{|\sqrt{n}(\mathbb{P}_n - P)(1(\tilde{u} \leq s))|}{s} \geq \frac{\sqrt{n} 2K \tilde{B}}{2B'} \right).$$

By Lemma 7.1, this is bounded above by a constant (that depends only on $\alpha_0, \beta_0, K, N_0$) times $1/rn$. Now, in terms of $\tilde{U}_1, \dots, \tilde{U}_n$ and $\tilde{\epsilon}_1, \dots, \tilde{\epsilon}_n$ (where the $\tilde{\epsilon}_i$'s are defined on the same probability space as the \tilde{U}_i 's, but independently of them, and are distributed like the ϵ_i 's) II_n is simply:

$$\Pr \left(\sup_{r/2K < s \leq \delta/2K} \frac{|(1/\sqrt{n}) \sum_{i=1}^n \tilde{\epsilon}_i (1(\tilde{U}_i \leq s))|}{s} \geq \frac{\sqrt{n} 2K \tilde{B}}{2} \right)$$

and this, by Lemma 7.3, is dominated up to a constant (that only depends on $\alpha_0, \beta_0, \sigma, K, N_0$) by $(1/rn)$. It follows that for some constant C_0 , for all $n \geq N_0$,

$$\sup_{(\alpha, \beta, t) \in R_n} P_{n1}(\alpha, \beta, t) \leq \frac{C_0}{rn}.$$

A similar (uniform) bound works $P_{n2}(\alpha, \beta, t)$. It follows that

$$\sup_{(\alpha, \beta, t) \in R_n} P_n(\alpha, \beta, t) \leq \frac{C_0}{rn}$$

at the expense of a larger constant C_0 . Thus, from (7.9), we have:

$$\sup_{(\alpha, \beta, t) \in R_n} \Pr_{\alpha, \beta, t}(|\hat{d}_{n2} - d^0| > rn^{-\gamma}) \leq C_0 (rn)^{-1} + \sup_{(\alpha, \beta, t) \in R_n} Q_n(\alpha, \beta, t).$$

To find a uniform upper bound on $Q_n(\alpha, \beta, t)$ note that, from (7.7), we have, for all $n > N_0$

$$\begin{aligned} \Pr_{\alpha, \beta, t}(|\hat{d}_{n_2} - d^0| > \delta n^{-\gamma}) &\leq \Pr_{\alpha, \beta, t}(\sup_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} |\tilde{M}_n(d) - \tilde{M}_n(d)| > b(\delta)) \\ &\leq \Pr_{\alpha, \beta, t}(\sup_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} |\tilde{M}_n(d) - \tilde{M}_n(d)| > \kappa \delta) \end{aligned}$$

and it suffices to find a uniform upper bound for this last expression. But this is bounded by

$$\begin{aligned} \Pr_{\alpha, \beta, t}[\sup_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} |\sqrt{n}(\tilde{\mathbb{P}}_n - \tilde{P}_n)(\mu(u) - (\alpha + \beta)/2)(1(u \leq d) - 1(u \leq d^0))| &> \sqrt{n}\kappa\delta/2] \\ &+ \Pr_{\alpha, \beta, t}[\sup_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} |n^{-1} \sum_{i=1}^n \epsilon_i(1(U_i \leq d) - 1(U_i \leq d^0))| > \kappa \delta/2]. \end{aligned}$$

To tackle the first term, we invoke Lemma 7.2. For $(\alpha, \beta, t) \in R_n$, the class $[(\mu(u) - (\alpha + \beta)/2)(1(u \leq d) - 1(u \leq d^0)) : d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]]$ is a bounded VC class of functions (with the bound not depending on α, β, t) and with finite VC dimension, say \mathcal{V} (which does not depend on α, β, t). Hence, we can apply Lemma 7.2 to conclude that:

$$\begin{aligned} \Pr_{\alpha, \beta, t}[\sup_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} |\sqrt{n}(\tilde{\mathbb{P}}_n - \tilde{P}_n)(\mu(u) - (\alpha + \beta)/2)(1(u \leq d) - 1(u \leq d^0))| &> \sqrt{n}\kappa\delta/2] \\ &\leq \tilde{C}_1 \times (\sqrt{n}\kappa\delta)^{2(\mathcal{V}-1)} \exp(-\tilde{C}_2 n \kappa^2 \delta^2), \end{aligned}$$

where the constants \tilde{C}_1 and \tilde{C}_2 depend solely on the VC dimension and the upper bound on the functions. For all sufficiently large n , the right side of the above display is less than $\epsilon/3$. To deal with the second term, we use the results on Pages 132–133 of Van de Geer (2000). We write the second term as:

$$\int \Pr_{\alpha, \beta, t}(\sup_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} |n^{-1} \sum_{i=1}^n \epsilon_i(1(u_i \leq d) - 1(u_i \leq d^0))| > \kappa \delta/2) dH_n(u_1, u_2, \dots, u_n)$$

where H_n is the joint distribution of (U_1, U_2, \dots, U_n) . For each fixed (u_1, u_2, \dots, u_n) , Corollary 8.8 of Van de Geer (2000) can be used to show that for δ sufficiently small and n sufficiently large (where the thresholds do not depend on the u_i 's or α, β, t),

$$\begin{aligned} \Pr_{\alpha, \beta, t}(\sup_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} |n^{-1} \sum_{i=1}^n \epsilon_i(1(u_i \leq d) - 1(u_i \leq d^0))| > \kappa \delta/2) \\ \leq \tilde{C} \exp(-\tilde{C}' n \delta^2), \end{aligned}$$

for some constants \tilde{C} and \tilde{C}' that do not depend on α, β, t or the points (u_1, u_2, \dots, u_n) . This implies that the second term can be made less than $\epsilon/3$ by choosing n sufficiently large. It follows, that for all sufficiently large n (say $n > N_1 > N_0$) and an appropriate choice of δ , we have:

$$\sup_{(\alpha, \beta, t) \in R_n} \Pr_{\alpha, \beta, t}(|\hat{d}_{n_2} - d^0| > r n^{-\gamma}) \leq C_0 (rn)^{-1} + 2\epsilon/3;$$

the first term on the right side can be made less than $\epsilon/3$ by choosing $r = A/n$ where A is large enough, showing that for all sufficiently large n , we can find A large enough so that:

$$\sup_{(\alpha, \beta, t) \in R_n} \Pr_{\alpha, \beta, t}(n^{1+\gamma} |\hat{d}_{n_2} - d^0| > A) < \epsilon.$$

It remains to say a few words about the application of Corollary 8.8 of Van de Geer (2000). The W_i s in that lemma are our ϵ_i s and we can, without loss, take $\mathcal{G} = \{g_d(u) = 1(u \leq d) - 1(u \leq d^0) : d \in [0, 1]\}$. The z_i 's are our u_i 's. The moment generating function condition on our errors implies that Condition (8.23) is satisfied for some σ_0^2 ; with $Q_n \equiv \sum_{i=1}^n \delta_{u_i}/n$, it is easy to see that the first condition of the corollary: $\sup_{d \in [0, 1]} \|g_d\|_{Q_n} \leq R$ is satisfied with $R = 1$. Condition (8.24) is satisfied with $K_2 = 1$. If δ is chosen to satisfy conditions (8.26) and (8.27), it is easy to see that Condition (8.28) will be automatically satisfied for sufficiently large n . This is based on the fact that $H_B(v, \mathcal{G}, Q_n) \leq \log(n+1)$ for $v > 1/\sqrt{n}$. To see this, construct brackets $\{[1(u \leq u_i) - 1(u \leq d^0), 1(u \leq u_{i+1}) - 1(u \leq d^0)]\}_{i=0}^n$, where $u_0 \equiv 0$ and $u_{n+1} \equiv 1$; any g_d lies in some bracket and the size of each of these brackets with respect to the Q_n norm is no larger than $1/\sqrt{n}$. Given a choice of δ , consider the right side of display (8.28) in Van de Geer (2000). This is given by

$$C_0 \int_{\delta/2^6}^{\sqrt{2}R\sigma_0} H_B^{1/2}(u/\sqrt{2}\sigma_0, \mathcal{G}, Q_n) du$$

and for all sufficiently large n , is readily seen to be dominated by

$$\sqrt{2}\sigma_0 C_0 \int_{1/\sqrt{n}}^R H_B^{1/2}(v, \mathcal{G}, Q_n) dv$$

which in turn is no larger than

$$\sqrt{2}\sigma_0 C_0 (R - 1/\sqrt{n}) \sqrt{\log(n+1)}.$$

But this quantity is certainly dominated by $\sqrt{n}\delta$ for large enough n . \square

Proof of Lemma 7.3: Let $\beta_n(s) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\epsilon}_i 1(x_i \leq s)$. Let $\{s_k = \alpha + (\beta - \alpha) 2^{-m} : 0 \leq k \leq 2^m\}$, $m \in \mathbb{N}$ be a dyadic partition of $[\alpha, \beta]$. Consider:

$$\begin{aligned}
P(\alpha, \beta, \lambda) &= P\left(\sup_{\alpha \leq s \leq \beta} \frac{|\beta_n(s)|}{s} \geq \lambda\right) \\
&= \lim_{m \rightarrow \infty} P\left(\max_{1 \leq k \leq 2^m} \frac{|\beta_n(s_k)|}{s_k} \geq \lambda\right) \\
&= \lim_{m \rightarrow \infty} \int_{(x_1, \dots, x_n) \in (0, 1)^n} P\left(\max_{1 \leq k < 2^m} \frac{\left|\frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\epsilon}_i 1(x_i \leq s_k)\right|}{s_k} \geq \lambda\right) \\
&\hspace{15em} dx_1 dx_2 \cdots dx_n
\end{aligned}$$

For fixed $(x_1, x_2, \dots, x_n) \in (0, 1)^n$, set

$$M_k = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\epsilon}_i 1(x_i \leq s_k), \quad 0 \leq k \leq 2^m.$$

Define $X_k = M_k - M_{k-1}$ for $k \geq 1$. Then the X_k 's are independent random variables, each with mean 0 and finite variance and $M_k = X_1 + X_2 + \dots + X_k$. Since $1/s_k$ is a decreasing sequence of constants, we can apply the Hajek-Renyi inequality (Lemma 7.4) to conclude that:

$$\begin{aligned}
P\left(\max_{1 \leq k \leq 2^m} \frac{|M_k|}{s_k} \geq \lambda\right) &\leq \frac{1}{\lambda^2} \sum_{1 \leq k \leq 2^m} s_k^{-2} E(M_k - M_{k-1})^2 \\
&= \frac{1}{\lambda^2} \sum_{1 \leq k \leq 2^m} s_k^{-2} (EM_k^2 - EM_{k-1}^2)
\end{aligned}$$

Now,

$$\begin{aligned}
EM_k^2 &= E\left[\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\epsilon}_i 1(x_i \leq s_k)\right)^2\right] \\
&= \text{Var}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\epsilon}_i 1(x_i \leq s_k)\right]^2 \\
&= \frac{1}{n} \sigma^2 \sum_{i=1}^n 1(x_i \leq s_k).
\end{aligned}$$

It follows that

$$EM_k^2 - EM_{k-1}^2 = \frac{\sigma^2}{n} \sum_{i=1}^n 1(s_{k-1} < x_i \leq s_k).$$

Therefore

$$\begin{aligned}
P(\alpha, \beta, \lambda) &\leq \lim_{m \rightarrow \infty} \int_{(x_1, \dots, x_n) \in (0,1)^n} \left\{ \frac{1}{\lambda^2} \sum_{1 \leq k < 2^m} s_k^{-2} \frac{\sigma^2}{n} \sum_{i=1}^n 1(s_{k-1} < x_i \leq s_k) \right\} \\
&\quad dx_1 dx_2 \cdots dx_n \\
&= \lim_{m \rightarrow \infty} \frac{1}{\lambda^2} \sum_{1 \leq k < 2^m} s_k^{-2} \frac{\sigma^2}{n} \sum_{i=1}^n \int 1(s_{k-1} < x_i \leq s_k) dx_i \\
&= \lim_{m \rightarrow \infty} \frac{1}{\lambda^2} \sigma^2 \sum_{1 \leq k < 2^m} s_k^{-2} (s_k - s_{k-1}) \\
&= \frac{\sigma^2}{\lambda^2} \int_{\alpha}^{\beta} \frac{1}{s^2} ds \\
&= (\alpha^{-1} - \beta^{-1}) \frac{\sigma^2}{\lambda^2}
\end{aligned}$$

□

References

- [1] Basseville, M. and Nikiforov, I. (1993), *Detection of Abrupt Changes: Theory and Application*, Prentice Hall, Englewood Cliffs, NJ.
- [2] Billingsley, P. (1999), *Convergence of Probability Measures*, Wiley Series in Probability and Statistics, New York, NY
- [3] Dempfle, A. and Stute, W. (2002), Nonparametric estimation of a discontinuity in regression, *Statistica Neerlandica*, 56, 233-242.
- [4] Ferger, D. (2004), A Continuous mapping theorem for the argmax-functional in the non-unique case, *Statistica Neerlandica*, 58, 83-96.
- [5] Ferger, D. (2005), On the minimizing point of the incorrectly centered empirical process and its limit distribution in nonregular experiments, *ESAIM: Probability and Statistics*, 9, 307-322.
- [6] Gijbels, I., Hall, P. and Kneip, A. (1999), On the estimation of jump points in smooth curves, *Annals of the Institute of Statistical Mathematics*, 51, 231-251.
- [7] Gregoire, G. and Hamrouni, Z. (2002), Change point estimation by local linear smoothing, *Journal of Multivariate Analysis*, 83, 56-83.
- [8] Hall, P. and Molchanov, I.S. (2003), Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces, *Annals of Statistics*, 31, 921-941.

- [9] Hung, Y.C. and Michailidis, G. (2007), Modeling, Analysis and Simulation of Switched Processing Systems, to appear in *ACM Transactions on Modeling, Analysis and Simulation*
- [10] Kosorok, M.R. (2006), *Introduction to Empirical Processes and Semiparametric Inference*, to appear in Springer Series in Statistics.
- [11] Kosorok, M.R. and Song, R. (2007), Inference under right censoring for transformation models with a change-point based on a covariate threshold, *Annals of Statistics*, 35, 957-989.
- [12] Lai, T.L. (2001), Sequential analysis: some classical problems and new challenges, *Statistica Sinica*, 11, 303-408
- [13] Loader, C.R. (1996), Change point estimation using nonparametric regression, *Annals of Statistics*, 24, 1667-1678.
- [14] Mueller, H.G. (1992), Change-points in nonparametric regression analysis, *Annals of Statistics*, 20, 737-761.
- [15] Mueller, H.G. and Song, K.S. (1997), Two-stage change-point estimators in smooth regression models, *Statistics and Probability Letters*, 34, 323-335.
- [16] Pons, O. (2003), Estimation in a Cox regression model with a change-point according to a threshold in a covariate, *Annals of Statistics*, 31, 442-463.
- [17] Ritov, Y. (1990), Asymptotic efficient estimation of the change point with unknown distributions, *The Annals of Statistics*, 18, 1829-1839.
- [18] Talagrand, M. (1994), Sharper bounds for Gaussian and empirical processes, *The Annals of Probability*, 22, 20-76.
- [19] Van de Geer, S. (2000), *Empirical Processes in M-Estimation*, Cambridge University Press, Cambridge, UK.