# TOPICS ON THRESHOLD ESTIMATION, MULTISTAGE METHODS AND RANDOM FIELDS

by

Atul Mallik

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2013

Doctoral Committee:

Associate Professor Moulinath Banerjee, Co-Chair
Professor Emeritus Michael B. Woodroofe, Co-Chair
Professor Alexander I. Barvinok
Professor Robert W. Keener
Professor George Michailidis

ACKNOWLEDGEMENTS

I take this opportunity to thank my advisors, Moulinath Banerjee and Michael Woodroofe, who have been nothing but kind and patient with me. I have never seen anybody with the energy and the enthusiasm that Mouli has for statistics, some of which rubbed off on me over the years. I am privileged to have worked with Michael who has been an inspiration to strive hard and stay humble. I consider it my good fortune to have them as my advisors. Emulating even a few of their qualities is an open research problem for me.

I would like to thank Professor Alexander Barvinok, Professor Robert Keener and Professor George Michailidis for agreeing to be on my committee. I am grateful to my collaborators George Michailidis and Bodhisattva Sen who have been wonderful people to work with. I would like to convey my sincere thanks to Jon Wellner, Wolfgang Polonik, Ambuj Tewari and Alexander Barvinok for finding time to meet and advise me on some of my projects. Finally, I would like to express my gratitude to the all the Statistics faculty members and especially, the staff who have helped me at some point or the other, during my stay at Michigan.

In this piece of space, I would like to thank my batchmates and the amazing number of friends I have made who have helped through my ups and downs. Naming a few would be an injustice to the rest. I am indebted to them in more than one way, in particular, for tolerating, what I call, my sense of humor.

I owe my deepest gratitude and respect to my parents and my brother. My mother has been a great teacher to me and a calming influence. My parents have

made uncountably many sacrifices to take care of my brother and me. My father had to quit his Ph.D. midway to come back to India and take care of his kids. Although he does not have any formal training in statistics, I am sure this thesis means a lot more to him than it can ever to me. I turn into a rude recluse at times but this has never stopped my brother from looking after me.

## TABLE OF CONTENTS

# LIST OF FIGURES

**Figure**

# LIST OF TABLES

# LIST OF APPENDICES

**Appendix**

# ABSTRACT

Topics on threshold estimation, multistage methods and random fields

by

Atul Mallik

Co-Chairs: Moulinath Banerjee and Michael Woodroofe

This dissertation addresses problems ranging from threshold estimation in Euclidean spaces to multistage procedures in M-estimation and central limit theorems for random fields.

We, first, consider the problem of identifying the threshold level at which a one-dimensional regression function leaves its baseline value. This is motivated by applications from dose-response studies and environmental statistics. We develop a novel approach that relies on the dichotomous behavior of $p$-value type statistics around this threshold. We study the large sample behavior of our estimate in two different sampling settings for constructing confidence intervals and also establish certain adaptive properties of our estimate.

The multi-dimensional version of the threshold estimation problem has connections to fMRI studies, edge detection and image processing. Here, interest centers on estimating a region (equivalently, its complement) where a function is at its baseline level. This is the region of no-signal (baseline region), which, in certain applications, corresponds to the background of an image; hence, identifying this region from noisy observations is equivalent to reconstructing the image. We study the computational

and theoretical aspects of an extension of the $p$-value procedure to this setting, primarily under a convex shape-constraint in two dimensions, and explore its applicability to other situations as well.

Multi-stage (designed) procedures, obtained by splitting the sampling budget suitably across stages, and designing the sampling at a particular stage based on information about the parameter obtained from previous stages, are often advantageous from the perspective of precise inference. We develop a generic framework for M-estimation in a multistage setting and apply empirical process techniques to develop limit theorems that describe the large sample behavior of the resulting M-estimates. Applications to change-point estimation, inverse isotonic regression, classification and mode estimation are provided: it is typically seen that the multistage procedure accentuates the efficiency of the M-estimates by accelerating the rate of convergence, relative to one-stage procedures. The step-by-step process induces dependence across stages and complicates the analysis in such problems, as careful conditioning arguments need to be employed for an accurate analysis.

Finally, in a departure from the more statistical components of the dissertation, we consider a central limit question for random fields. Random fields – real valued stochastic processes indexed by a multi-dimensional set – arise naturally in spatial data analysis and image detection. Limit theorems for random fields have, therefore, received considerable interest. We prove a Central Limit Theorem (CLT) for linear random fields that allows sums to be taken over sets as general as the disjoint union of rectangles. A simple version of our result provides a complete analogue of a CLT for linear processes with a lot of uniformity, at the expense of no extra assumptions.

# CHAPTER 1

# Introduction

A principal task in modern non-parametrics is to devise methods to solve *non-standard* problems, problems in which the convergence rate of estimates is different from $\sqrt{n}$. This is typically due to the parameter of interest being at the boundary or the non-smooth nature of the model. These non-standard problems find applications in a variety of disciplines such as genomics, astrophysics, finance, pharmacology, environmental statistics, image processing and other related fields. Coming up with efficient estimates and studying their properties is a major challenge in these settings. It requires applying and extending results from empirical process theory, especially for M-estimation methods, which forms a major portion of this dissertation. One such non-standard problem of threshold estimation is studied in Chapter 2 of this dissertation, extensions and variants of which are considered in detail over Chapters 3 to 5. In Chapter 6, we provide a general treatment of multi-stage procedures which are found useful in several non-standard problems such as change-point estimation, inverse isotonic regression and mode estimation.

In a departure from problems with statistical flavor, a part of this dissertation addresses a central limit question from applied probability. Central limit theorems (CLTs) answer how (normalized) partial sums of random variables behave asymptotically in a variety of settings. They are the cornerstone for doing large sample inference in statistics and pose exciting challenges to the theorists. Recently, extending CLTs for random variables to that for random fields has received considerable interest as random fields – real valued stochastic processes indexed by a multi-dimensional set – arise quite naturally in applications from spatial data analysis, statistical mechanics and image processing. We approach one such problem in Chapter 7 of this dissertation.

This thesis covers three broad topics based on a series of papers and articles by us. We provide a summary of these topics in the following section.

## 1.1 Summary and organization of the thesis

### 1.1.1 Part I: Threshold estimation in various settings

Over the course of Chapters 2 to 5, we study the problem of identifying the threshold level (equivalently, a region in higher dimensions) at which a regression function leaves its baseline value. This problem is motivated by applications that arise in toxicological and pharmacological dose-response studies, environmental statistics, engineering, image processing and other related fields. In the one-dimensional setting, we consider a data generating model of the form $Y = \mu(X) + \epsilon$, where $\mu : [0,1] \mapsto \mathbb{R}$ satisfies the property that $\mu(x) = \tau_0$ for $x \leq d_0$ and $\mu(x) > \tau_0$ for $x > d_0$ for unknown $\tau_0$ and $d_0$. The interest centers around estimating the threshold level $d_0$.

In Chapter 2, we come up with a novel approach for estimating $d_0$ that relies on the dichotomous behavior of certain $p$-values on either side of $d_0$. We study the procedure for two different sampling settings, one where several responses can be obtained at a number of different covariate-levels *(dose-response)* and the other involving limited number of response values per covariate *(standard regression)*. The estimate is shown to be consistent and its finite sample properties are studied extensively through simulations. Our approach is computationally simple and extends to the estimation of the baseline value $\tau$ of the regression function, situations with heteroscedastic errors and to time-series. We illustrate our approach on some real data applications.

This part of the thesis is based on joint work with Bodhisattva Sen, Moulinath Banerjee and George Michailidis. It appears in our paper Mallik et al. (2011).

In Chapters 3 and 4, we further delve into the large sample properties of our $p$-value based estimate in the dose–response and the standard regression settings. The two settings require fairly different treatment and yield markedly different limit distributions. However, they exhibit the same rate of convergence. The smoothness of the regression function in the vicinity of the threshold plays an important role in determining the rate of convergence. A "cusp" of order $k$ at the threshold $d_0$ yields an optimal rate of $N^{-1/(2k+1)}$, where $N$ is the total budget. In Chapter 3, we apply non-standard empirical process techniques such as argmin continuous mapping in the non-unique case to show that the estimate of $d_0$ in the dose–response setting converges to a minimizer of a generalized compound Poisson process. Based on the limiting behavior, we provide a recipe for constructing confidence intervals which we study through a limited simulation study and apply to a dataset from a complex queuing system.

The estimate for $d_0$ in the standard regression setting is constructed via kernel estimators which, in spite of starting with independent observations, induce dependence. We address this and other intricacies of the standard regression setting in Chapter 4, where in conjunction with standard empirical process techniques (meant for independent and identically distributed random variables), we apply blocking arguments and martingale inequalities to deduce the rate of convergence. We show that the asymptotic distribution of the normalized estimate of the threshold is the minimizer of an integrated and transformed Gaussian process. We study the finite sample behavior of confidence intervals obtained through the asymptotic approximation using simulations, consider extensions to short-range dependent data, and apply our inference procedure to two datasets from Chapter 3.

Chapters 3 and 4 include collaborative work with Bodhisattva Sen, Moulinath Banerjee and George Michailidis. They appear in our articles Mallik et al. (2013a) and Mallik et al. (2013b).

In Chapter 5, we consider the multi-dimensional version of threshold estimation problem, i.e., we now have a regression function $\mu : \mathbb{R}^d \mapsto \mathbb{R}$ such that $\mu(x) = \tau_0$ for $x \in S_0$ and $\mu(x) > \tau_0$ for $x \in S_0^c$. The problem of identifying the baseline region $S_0$ (equivalently, its complement) arises in a broad range of problems, e.g., determining high pollution zones in a densely inhabited region, finding active regions from fMRI studies, image processing and edge detection. The set $S_0$ is the region of no-signal (baseline region), which, in certain applications, corresponds to the background of an image; hence, identifying this region from noisy observations is equivalent to reconstructing the image. The $p$-value procedure for the one-dimensional case has a natural extension to this setting but the computational and the theoretical aspects of the problem become more involved due to complex nature of the set $S_0$, no longer being identified by a single point $d_0$, and the local behavior of the function $\mu$ near the boundary of $S_0$. We primarily consider the case with a convex shape-constraint on $S_0$ and a cusp type assumption on $\mu$ at the boundary of $S_0$. We explore the applicability of our approach to other situations as well.

This part of the thesis is based on joint work with Moulinath Banerjee and Michael Woodroofe.

### 1.1.2 Part II: Multistage procedures

Multi-stage procedures involve splitting the sampling budget suitably across stages and typically involve designing the sampling at a particular stage based on information about the parameter obtained from previous stages. They are often found advantageous from the perspective of precise inference in various non-parametric settings as they typically accentuates the efficiency of the estimates by accelerating the rate of convergence, relative to one-stage procedures. However, the step-by-step process induces dependence across stages and complicates the analysis in such problems. In Chapter 6, we develop a generic framework for M-estimation in a multistage setting. We apply empirical process techniques and careful conditioning arguments to develop limit theorems that describe the large sample behavior of the resulting M-estimates. This unified approach is illustrated on a variety of problems ranging from change-point estimation to inverse isotonic regression and mode estimation.

This part of the thesis is based on joint work with Moulinath Banerjee and George Michailidis.

### 1.1.3 Part III: Random fields

As mentioned earlier, random fields are real valued stochastic processes indexed by a multi-dimensional set which arise naturally in spatial data analysis, image detection and related fields. We are mainly concerned with proving Central Limit Theorems (CLT) for *linear random fields* where sums are taken over sets of arbitrary shape. Most approaches in the literature rely upon the use of *Beveridge–Nelson decomposition* to derive conditions for CLT when sums are taken over rectangles. In Chapter 7, we provide a different approach that extends the CLT in Ibragimov (1962) for linear processes to that for linear random fields without putting any extra assumptions. In its most general form, we prove a CLT when sums of linear random fields are considered over disjoint union of rectangles.

This part of the thesis is based on joint work with Michael Woodroofe. It appears in our paper Mallik and Woodroofe (2011).

# Part I

# Threshold estimation in various settings

# CHAPTER 2

# Threshold Estimation using $p$-values

In a number of applications, the data follow a regression model where the regression function $\mu$ is constant at its baseline value $\tau_0$ up to a certain covariate threshold $d_0$ and deviates significantly from $\tau_0$ at higher covariate levels. For example, consider the data shown in the left panel of Fig. 2.1. It depicts the physiological response of cells from the IPC-81 leukemia rat cell line to a treatment, at different doses; more details are given in Section 2.2.5. The objective here is to study the toxicity in the cell culture to assess environmental hazards. The function stays at its baseline value for high dose levels which corresponds to the dose becoming lethal, and then takes off for lower doses, showing response to treatment. This problem requires procedures that can identify the change-point in the regression function, namely where it deviates from the baseline value. The threshold is of interest as it corresponds to maximum safe dose level beyond which cell cultures stop responding. Similar problems also arise in other toxicological applications (Cox, 1987).

Problems with similar structure also arise in other pharmacological dose-response studies, where $\mu(x)$ quantifies the response at dose-level $x$ and is typically at the baseline value up to a certain dose, known as the minimum effective dose; see Chen and Chang (2007) and Tamhane and Logan (2002) and the references therein. In such applications, the number of doses or covariate levels is relatively small, say up to 20,

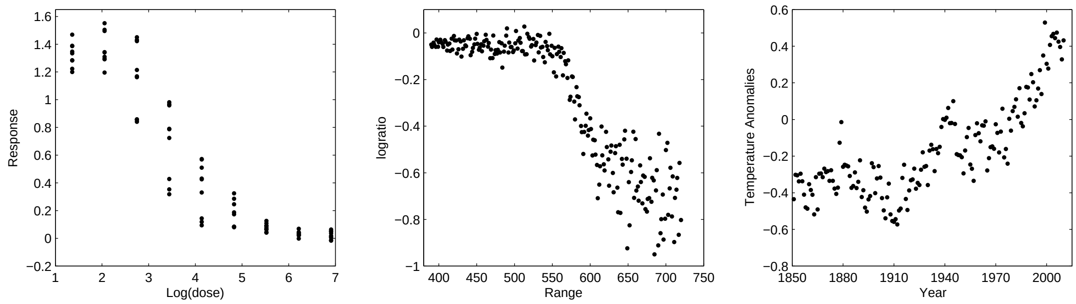



Figure 2.1: The three data examples. Left panel: Response of cell-cultures at different doses. Middle panel: Logratio measurements over range. Right panel: Annual global temperature anomalies from 1850 to 2009.

and many procedures proposed in the literature are based on testing ideas (Tamhane and Logan, 2002; Hsu and Berger, 1999). However, in other application domains, the number of doses can be fairly large compared to the number of replicates at each dose. The latter is effectively the setting of a standard regression model. In the extreme case, there is a single observation per covariate level. Data from such a setting are shown in the middle panel of Fig. 2.1, depicting the outcome of a light detection and ranging (LIDAR) experiment, used to detect the change in the level of atmospheric pollutants. This technique uses the reflection of laser-emitted light to detect chemical compounds in the atmosphere (Holst et al., 1996; Ruppert et al., 1997). The predictor variable, range, is the distance traveled before the light is reflected back to its source, while the response variable, logratio, is the logarithm of the ratio of received light at two different frequencies. The negative of the slope of the underlying regression function is proportional to mercury concentration at any given value of range. The point at which the function falls from its baseline level corresponds to an emission plume containing mercury and, thus, is of interest. An important difference between these two examples is that the former provides the luxury of multiple observations at each covariate level, while the latter does not.

Another relevant application in a time-series context is given in the right panel of

Fig. 2.1, where annual global temperature anomalies are reported from 1850 to 2009. The study of such anomalies, temperature deviations from a base value, has received much attention in the context of global warming from both the scientific as well as the general community (Melillo, 1999; Delworth and Knutson, 2000). The figure suggests an initial flat stretch followed by a rise in the function. Detecting the advent of global warming, which is the threshold, is of interest here. While we take advantage of the independence of errors in the previous two datasets, this application has an additional feature of short range dependence which needs to be addressed appropriately.

Formally, we consider a function $\mu(x)$ on $[0,1]$ with the property that $\mu(x) = \tau_0$ for $x \leq d_0$ and $\mu(x) > \tau_0$ for $x > d_0$ for some $d_0 \in (0,1)$. As already mentioned, quantities of prime interest are $d_0$ and $\tau_0$ that need to be estimated from realizations of the model $Y = \mu(X) + \epsilon$. We call $d_0$ the $\tau_0$ threshold of the function $\mu$. Here $\tau_0$ is the global minimum for the function $\mu$. To fix ideas, we work only with this setting in mind. The methods proposed can be easily imitated for the first data application where the baseline stretch is on the right as well as for the second data application where $\tau_0$ is the maximum.

In this generality, i.e., without any assumptions on the behavior of the function in a neighborhood of $d_0$, the estimation of the threshold $d_0$ has not been extensively addressed in the literature. In the simplest possible setting of the problem posited, $\mu$ has a jump discontinuity at $d_0$. In this case, $d_0$ corresponds to a change-point for $\mu$ and the problem reduces to estimating this change-point. Such models are well studied; see Loader (1996), Koul et al. (2003), Pons (2003), Lan et al. (2009) and the references therein. Results on estimating a change-point in a density can be found in Ibragimov and Has′minskiĭ (1981).

The problem becomes significantly harder when $\mu$ is continuous at $d_0$; in particular, the smoother $\mu$ is in a neighborhood of $d_0$, the more challenging the estimation. If $d_0$ is a cusp of $\mu$ of some *known* order $k$, i.e., the first $k-1$ right derivatives of $\mu$ at $d_0$

equal 0 but the $k$-th does not, so that $d_0$ is a change-point in the $k$-th derivative, one can obtain nonparametric estimates for $d_0$ using either kernel based (Müller, 1992) or wavelet based (Raimondo, 1998) methods. If the degree of differentiability of $\mu$ at $d_0$ *is not known*, this becomes an even harder problem.

In this chapter, we develops a novel approach for the *consistent estimation* of $d_0$ in situations where *single or multiple observations can be sampled* at a given covariate value. The developed nonparametric methodology relies on testing for the value of $\mu$ at the design values of the covariate. The obtained test statistics are then used to construct $p$-values which, under mild assumptions on $\mu$, behave in markedly different manner on either side of the threshold $d_0$ and it is this discrepancy that is used to construct an estimate of $d_0$. The approach is computationally simple to implement and does not require knowledge of the smoothness of $\mu$ at $d_0$. In a dose-response setting involving several doses and large number of replicates per dose, the $p$-values are constructed using multiple observations at each dose. The approach is completely automated and does not require the selection of any tuning parameter. In the case of limited or even single observation at each covariate value, referred to as the standard regression setting, the $p$-values are constructed by borrowing information from neighboring covariate values via smoothing which only involves selecting a smoothing bandwidth. The first data application falls under the dose-response setting and the other two examples fall under the standard regression regime. We establish consistency of the proposed procedure in both settings.

An estimate of $\mu$, say $\hat{\mu}$, by itself, fails to offer a satisfactory solution for estimating $d_0$. Naive estimates, using $\hat{\mu}$, may be of the form $\hat{d}^{(1)} = \sup\{x : \hat{\mu}(x) \leq \tau_0\}$ or $\hat{d}^{(2)} = \inf\{x : \hat{\mu}(x) > \tau_0\}$. The estimator $\hat{d}^{(1)}$ performs poorly when $\mu$ is not monotone, and is close to $\tau_0$ at values to the far right of $d_0$, e.g., when $\mu$ is tent-shaped. Also, $\hat{d}^{(2)}$, by itself, is not consistent and one would typically need to substitute $\tau_0$ with a $\tau_0 + \eta_n$, with $\eta_n \to 0$ at an appropriate rate, to attain consistency. In contrast, our approach does not need to introduce such exogenous parameters.

## 2.1 Formulation and Methodology

### 2.1.1 Problem Formulation

Consider a regression model $Y = \mu(X) + \epsilon$, where $\mu$ is a function on $[0, 1]$ and

$$\mu(x) = \tau_0 \text{ for } x \leq d_0, \text{ and } \mu(x) > \tau_0 \text{ for } x > d_0, \tag{2.1}$$

for $d_0 \in (0, 1)$, with an unknown $\tau_0 \in \mathbb{R}$. The covariate $X$ is sampled from a Lebesgue density $f$ on $[0, 1]$ and $E(\epsilon \mid X = x) = 0$, $\sigma^2(x) = \text{var}(\epsilon \mid X = x) > 0$ for $x \in [0, 1]$. We assume that $f$ is continuous and positive on $[0, 1]$ and $\mu$ is continuous. *No further assumptions are made on the behavior of $\mu$, especially around $d_0$.* We have the following realizations:

$$Y_{ij} = \mu(X_i) + \epsilon_{ij}, \; i = 1, \ldots, n; \; j = 1, \ldots, m, \tag{2.2}$$

with $N = m \times n$ being the total budget of samples. The $\epsilon_{ij}$s are independent given $X$ and distributed like $\epsilon$ and the $X_i$s are independent realizations from $f$. Also, (2.2) with $m = 1$ corresponds to the usual regression setting which simply has only one response at each covariate level.

We construct *consistent* estimates of $d_0$ in dose-response and standard regression settings. In the dose-response setting, we allow both $m$ and $n$ to be large and construct $p$-values accordingly. We refer to the corresponding approach as Method 1 from now on. In the other setting, we consider the case when $m$ is much smaller compared to $n$ and extend our approach through smoothing. We refer to this extension as Method 2, which requires choosing a smoothing bandwidth. The two methods rely on the same dichotomous behavior exhibited by the approximate $p$-values, although constructed differently.

### 2.1.2 Dose-Response Setting (Method 1)

We start by introducing some notation. Let $\bar{Y}_{i\cdot} = \sum_{i=1}^{m} Y_{ij}/m$ and $x \in (0,1)$ denote a generic value of the covariate. Let $\hat{\sigma}_{m,n} \equiv \hat{\sigma}$ and $\hat{\tau}_{m,n} \equiv \hat{\tau}$ denote the estimators of $\sigma(\cdot)$ and $\tau_0$ respectively. For homoscedastic errors, $\hat{\sigma}_{m,n}(\cdot)$ is the standard pooled estimate, i.e., $\hat{\sigma}^2_{m,n}(x) \equiv \sum_{i,j}(Y_{ij} - \bar{Y}_{i\cdot})^2/(nm-m)$, while for the heteroscedastic case $\hat{\sigma}^2_{m,n}(X_i) = \sum_{j=1}^{m}(Y_{ij} - \bar{Y}_{i\cdot})^2/(m-1)$. Estimators of $\tau_0$ are discussed in Section 2.1.4. We seek to estimate $d_0$ by constructing $p$-values for testing the null hypothesis $H_{0,x} : \mu(x) = \tau_0$ against the alternative $H_{1,x} : \mu(x) > \tau_0$ at each dose $X_i = x$. The approximate $p$-values are

$$p_{m,n}(X_i) = p_{m,n}(X_i, \hat{\tau}_{m,n}) = 1 - \Phi(\sqrt{m}(\bar{Y}_{i\cdot} - \hat{\tau})/\hat{\sigma}(X_i)).$$

Indeed, these approximate $p$-values would correspond to the exact $p$-values for the uniformly most powerful test if we worked with a known $\sigma$, a known $\tau$ and normal errors.

To the left of $d_0$, the null hypothesis holds and these approximate $p$-values converge weakly to a Uniform(0,1) distribution, for suitable estimators of $\tau_0$. In fact, the distribution of $p_{m,n}(X_i)$s does not even depend on $X_i$ when $X_i \leq d_0$. Moreover, to the right of $d_0$, where the alternative is true, the $p$-values converge in probability to 0. This dichotomous behavior of the $p$-values on either side of $d_0$ can be used to prescribe *consistent* estimates of the latter. We can fit a stump, a piecewise constant function with a single jump discontinuity, to the $p_{m,n}(X_i)$s, $i = 1, \ldots, n$, with levels 1/2, which is the mean of a Uniform (0,1) random variable, and 0 on either side of the break-point and prescribe the break-point of the best fitting stump (in the sense of least squares) as an estimate of $d_0$. Formally, we fit a stump of the form

$\xi_d(x) = (1/2)1(x \leq d)$, minimizing

$$\tilde{\mathbb{M}}_{m,n}(d) = \tilde{\mathbb{M}}_{m,n}(d, \hat{\tau}) = \sum_{i:X_i \leq d} \{p_{m,n}(X_i) - 1/2\}^2 + \sum_{i:X_i > d} \{p_{m,n}(X_i)\}^2 \tag{2.3}$$

over $d \in [0, 1]$. Let $\hat{d}_{m,n} = \operatorname{argmin}_{d \in [0,1]} \tilde{\mathbb{M}}_{m,n}(d)$. The success of our method relies on the fact that the $p_{m,n}(X_i)$s eventually show stump like dichotomous behavior. In this context, no estimate of $\mu$ could exhibit such a behavior directly. Our procedure can be thought of as fitting the limiting stump model to the observed $p_{m,n}(X_i)$s by minimizing an $L_2$ norm. In fact, the expression in (2.3) can be simplified. Let

$$\mathbb{M}_{m,n}(d) = \frac{1}{n} \sum_{i:X_i \leq d} \left(p_{m,n}(X_i) - (1/4)\right).$$

Elementary calculations show that

$$\hat{d}_{m,n} = \underset{d \in [0,1]}{\operatorname{sargmax}} \, \mathbb{M}_{m,n}(d).$$

Here, sargmax denotes the smallest argmax of the criterion function, which does not have a unique maximum. In fact, $\hat{d}_{m,n}$ corresponds to an order statistic of $X_i$s and the above criterion is maximized at any point between $\hat{d}_{m,n}$ and the next order statistic. Our results hold for any maximizer of the criterion; the smallest argmax is chosen just to fix ideas. The estimate is easy to compute as it requires a simple search over the order statistics.

In heteroscedastic models, the estimation of the error variance $\hat{\sigma}(\cdot)$ can often be tricky. The proposed procedure can be modified to *avoid* the estimation of the error variance altogether for the construction of the $p$-values, as the desired dichotomous behavior of the $p$-values is preserved even when we do not normalize by the estimate of the variance. Thus, we can consider the modified $p$-values $\tilde{p}_{m,n}(X_i) = 1 - \Phi(\sqrt{m}(\bar{Y}_{i\cdot} - \hat{\tau}))$ and the dichotomy continues to be preserved as

$E\{1 - \Phi(Z)\} = 0.5$ for a normally distributed $Z$ with zero mean and arbitrary variance. In practice though, we recommend, whenever possible, using the normalized $p$-values as they exhibit good finite sample performance.

Next, we prove the consistency of our proposed procedure when using the unnormalized $p$-values. The technique illustrated here can be carried forward to prove consistency for other variants of the procedure, e.g., when normalizing by the estimate of the error variance, but require individual attention depending upon the assumption of heteroscedasticity/homoscedasticity.

**Theorem 2.1.** *Consider the dose-response setting of the problem and let $\hat{d}_{m,n}$ denote the estimator based on the non-normalized version of p-values, e.g., $\tilde{p}_{m,n}(X_i) = 1 - \Phi(\sqrt{m}(\bar{Y}_{i\cdot} - \hat{\tau}))$. Assume that $\sqrt{m}(\hat{\tau}_{m,n} - \tau_0) = o_p(1)$ as $m, n \to \infty$, i.e., given $\epsilon, \eta > 0$, there exists a positive integer $L$, such that for $m, n \geq L$, $P(\sqrt{m}|\hat{\tau} - \tau_0| > \epsilon) < \eta$. Then, $\hat{d}_{m,n} - d_0 = o_p(1)$ as $m, n \to \infty$.*

The proof is relegated to Section A.1 of Appendix A.

### 2.1.3 Standard Regression Setting (Method 2)

We now consider the case when $m$ is much smaller than $n$. Let $\hat{\mu}(x) = \hat{r}(x)/\hat{f}(x)$ denote the Nadaraya–Watson estimator, where

$$\begin{aligned}
\hat{r}(x) &= \frac{1}{nh_n}\sum_{i=1}^{n} \bar{Y}_{i\cdot} K\left(\frac{x - X_i}{h_n}\right) \text{ and} \\
\hat{f}(x) &= \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right),
\end{aligned}$$

with $K$ being a symmetric probability density (a kernel) and $h_n$ the smoothing bandwidth. We take $h_n = cn^{-\beta}$ for $\beta \in (0, 1)$. Let $\hat{\sigma}_n(\cdot)$ and $\hat{\tau}_n$ denote estimators of $\sigma(\cdot)$ and $\tau_0$ respectively. An estimate of $\sigma^2(\cdot)$ can be constructed through standard techniques, e.g., smoothing or averaging the squared residuals $m(\bar{Y}_{i\cdot} - \hat{\mu}(X_i))^2$, depending upon the assumption of heteroscedastic or homoscedastic errors.

For $x < d_0$, the statistic $T(x, \tau_0) = \sqrt{nh_n}(\hat{\mu}(x) - \tau_0)$ converges to a normal distribution with zero mean and variance $\Sigma^2(x) = \Sigma^2(x, \sigma) = \sigma^2(x)\bar{K}^2/\{mf(x)\}$ with $\bar{K}^2 = \int K^2(u)du$. The approximate $p$-value for testing $H_{0,x}$ against $H_{1,x}$ can then be constructed as:

$$p_n(x) = p_n(x, \hat{\tau}_n) = 1 - \Phi\left(T(x, \hat{\tau}_n)/\Sigma(x, \hat{\sigma})\right),$$

where $\Sigma^2(x, \hat{\sigma}) = \hat{\sigma}_n^2(x)\bar{K}^2/\{m\hat{f}(x)\}$. It can be seen that these $p$-values also exhibit the desired dichotomous behavior. Finally, an estimate of $d_0$ is obtained by maximizing

$$\mathbb{M}_n(d) = \frac{1}{n}\sum_{i:X_i \le d}\{p_n(X_i) - 1/4\} \tag{2.4}$$

over $d \in [0, 1]$. Let $\hat{d}_n = \operatorname{sargmax}_{d\in[0,1]} \mathbb{M}_n(d)$. Under suitable conditions on $\hat{\tau}_n$, this estimator can be shown to be consistent when $n$ grows large.

We have avoided sophisticated means of estimating $\mu(\cdot)$, as our focus is on estimation of $d_0$, and not particularly on efficient estimation of the regression function. Also, the Nadaraya–Watson estimate does not add substantially to the computational complexity of the problem and provides a reasonably rich class of estimators through choices of bandwidths and kernels.

In many applications, particularly with $m = 1$ and heteroscedastic errors, estimating the variance function $\sigma^2(\cdot)$ accurately could be cumbersome. As with Method 1, Method 2 can also be modified to avoid estimating the error variance, e.g., the estimator constructed using (2.4), based on $\tilde{p}_n(X_i)$s, with $\tilde{p}_n(x) = 1 - \Phi\left(\sqrt{nh_n}(\hat{\mu}(x) - \hat{\tau}_n)\right)$. Next, we prove consistency for the proposed procedure *when we do not normalize by the estimate of the variance.* The technique illustrated here can be carried forward to prove consistency for other variants of the procedure. We make the following additional assumptions.

(a) For some $\eta > 0$, the functions $\sigma^2(\cdot)$ and $\sigma^{(2+\eta)}(x) \equiv E(|\epsilon|^{2+\eta} \mid X = x)$, $x \in [0, 1]$, are continuous.

(b) The kernel $K$ is either compactly supported or has exponentially decaying tails, i.e., for some $C$, $D$ and $a > 0$, and for all sufficiently large $x$, $P\{|W| > x\} \leq C\exp(-Dx^a)$, where $W$ has density $K$. Also, $\bar{K^2} = \int K^2(u)du < \infty$.

Assumption (a) is very common in non-parametric regression settings for justifying asymptotic normality of kernel based estimators. Also, the popularly used kernels, namely uniform, Gaussian and Epanechnikov, do satisfy assumption (b).

**Theorem 2.2.** *Consider the standard regression setting of the problem with $m$ staying fixed and $n \to \infty$. Assume that $\sqrt{nh_n}(\hat{\tau}_n - \tau_0) = o_p(1)$ as $n \to \infty$. Let $\hat{d}_n$ denote the estimator computed using $\tilde{p}_n(X_i) = 1 - \Phi\{T(X_i, \hat{\tau}_n)\}$. Then, $\hat{d}_n - d_0 = o_p(1)$ as $n \to \infty$.*

The proof is given in Section A.2 of Appendix A.

**Remark 2.3.** *The model in (2.2) incorporates situations with discrete responses. For example, we can consider binary responses with $Y_{ij}$s indicating a reaction to a dose at level $X_i$. We assume that the function $\mu(x)$, the probability that a subject yields a reaction at dose $x$, is of the form (2.1) and takes values in $(0, 1)$ so that $\sigma^2(x) = \mu(x)(1 - \mu(x)) > 0$. The results from this section as well as those from Section 2.1.2 will continue to hold for this setting.*

**Remark 2.4.** *Our assumption of continuity of $\mu$ can be dropped and the results from this section as well as those from Section 2.1.2 will continue to hold provided that $\mu$ is bounded and continuous almost everywhere with respect to Lebesgue measure. This includes the classical change-point problem where $\mu$ has a jump discontinuity at $d_0$ but is otherwise continuous.*

### 2.1.4 Estimators of $\tau_0$

Suitable estimates of $\tau_0$ are required that satisfy the conditions stated in Theorems 2.1 and 2.2. In a situation where $d_0$ may be safely assumed to be greater than some known positive $\eta$, an estimate of $\tau_0$ can be obtained by taking the average of the response values on the interval $[0, \eta]$. The estimator would be $\sqrt{mn}$-consistent and would therefore satisfy the required conditions. Such an estimator is seen to be reasonable for most of the data applications that are considered in this chapter. In situations when such a solution is not satisfactory, we propose an approach to estimate $\tau_0$ that does not require any background knowledge, once again using $p$-values.

We now construct an explicit estimator $\hat{\tau}$ of $\tau_0$ in the dose-response setting, as required in Theorem 2.1, using $p$-values. For convenience, let

$$Z_{im}(\tau) = p_{m,n}(X_i, \tau) = 1 - \Phi\left(\sqrt{m}(\bar{Y}_{i\cdot} - \tau)/\hat{\sigma}_{m,n}(X_i)\right).$$

Let $\tau > \tau_0$. As $m$ increases, for $\mu(X_i) < \tau$, $Z_{im}(\tau)$ converges to 1 in probability, while for $\mu(X_i) > \tau$, $Z_{im}(\tau)$ converges to 0 in probability. For any $\tau < \tau_0$, it is easy to see that $Z_{im}(\tau)$ always converges to 0, whereas when $\tau = \tau_0$, $Z_{im}(\tau)$ converges to 0 for $X_i > d_0$ and $E\{Z_{im}(\tau)\}$ converges to 1/2 for $X_i < d_0$. Thus, it is only when $\tau = \tau_0$ that $Z_{im}(\tau)$s are closest to 1/2 for a substantial number of observations. This suggests a natural estimate of $\tau_0$:

$$\hat{\tau} \equiv \hat{\tau}_{m,n} = \underset{\tau}{\operatorname{argmin}} \sum_{i=1}^{n} \{Z_{im}(\tau) - 1/2\}^2. \tag{2.5}$$

Theorem 2.5 shows that under some mild conditions and homoscedasticity, $\sqrt{m}\,(\hat{\tau}_{m,n} - \tau_0)$ is $o_p(1)$, a condition required for Theorem 2.1.

**Theorem 2.5.** *Consider the same setup as in Theorem 2.1. Assume that the errors are homoscedastic with variance $\sigma_0^2$. Further suppose that the regression function $\mu$*

*satisfies:*

*(A) Given $\eta > 0$, there exists $\epsilon > 0$ such that, for every $\tau > \tau_0$,*

$$\int_{\{x>d_0:|\mu(x)-\tau|\leq\epsilon\}} f(x)dx < \eta.$$

*Also assume that $\phi_m$, the density function of $\sqrt{m}\,\bar{\epsilon}_{1.}/\sigma_0$, converges pointwise to $\phi$, the standard normal density. Then $\sqrt{m}\,(\hat{\tau}_{m,n} - \tau_0) = o_p(1)$.*

This proof is given in Section A.3 of Appendix A.

**Remark 2.6.** *Condition (A) is guaranteed if, for example, $\mu$ is strictly increasing to the right of $d_0$ although it holds under weaker assumptions on $\mu$. In particular, it rules out flat stretches to the right of $d_0$. The assumption that $\phi_m$ converges to $\phi$ is not artificial, since convergence of the corresponding distribution functions to the distribution function of the standard normal is guaranteed by the central limit theorem.*

This approach in (2.5) can also be emulated to construct estimators of $\tau_0$ for the standard regression setting by just going through the procedure with $p_n(X_i, \tau)$s instead of $p_{m,n}(X_i, \tau)$s and it is clear that this estimator is consistent. However, the theoretical properties of this estimator, such as the rate of convergence, are not completely known. Nevertheless, the procedure has good finite sample performance as indicated by the simulation studies in Section 2.2. The estimator is positively biased. This is due to the fact that a value larger than $\tau_0$ is likely to minimize the objective function in (2.5) as it can possibly fit the $p$-values arising from a stretch extending beyond $[0, d_0]$, in presence of noisy observations. The values smaller than $\tau_0$ do not get such preference as the true function never falls below $\tau_0$.

### 2.1.5 To smooth or not to smooth

The consistency of the two methods established in the previous sections justifies good large sample performance of the procedures, but does not provide us with practical guidelines on which method to use given a real application. In dose-response studies, it is quite difficult to find situations where both $m$ and $n$ are large. Typically, such studies do not administer too many dose levels which precludes $n$ from being large. So, we compare the finite sample performance of the two methods for different allocations of $m$ and $n$ to highlight their relative merits.

We study the performance of the two methods for three different choices of regression functions. All these functions are assumed to be at the baseline value 0 to the left of $d_0 \equiv 0.5$. Specifically, $M_1$ is a piece-wise linear function rising from 0 to 0.5 between $d_0$ and 1; $M_2$, a convex curve, grows like a quadratic beyond $d_0$, and reaches 0.5 at 1; $M_3$ rises linearly with unit slope for values ranging from $d_0$ to 0.8 and then decreases with unit slope for values between 0.8 and 1.0. So, $M_1$ and $M_2$ are strictly monotone to the right of $d_0$ and exhibit increasing level of smoothness at $d_0$. On the other hand, $M_3$ is tent-shaped and estimating $d_0$ is expected to be harder for $M_3$ compared to $M_1$.

For each allocation pair $(m, n)$ and a choice of a regression function, we generate responses $\{Y_{i1}, \ldots, Y_{im}\}$, with $Y_{ij} = \mu(X_i) + \epsilon_{ij}$, the $\epsilon_{ij}$s being independent $N(0, \sigma^2)$ with $\sigma = 0.3$. The $X_i$s are sampled from Uniform(0,1). The performance for estimating $d_0 \equiv 0.5$ is studied based on root mean square error computed over 2000 replicates, assuming a known variance and a known $\tau_0 \equiv 0$. For illustrative purposes, we use the Gaussian kernel for Method 2. It will be seen that a bandwidth of the form $h_n = h_0 n^{-1/(2k+1)}$ is chosen as it is expected to attain the optimal rate of convergence for estimating a cusp of order $k$ (see also Raimondo (1998)). For $M_1$ and $M_3$, $k = 1$ while for $M_2$, $k$ is 2. We report the simulations for the best $h_0$ which minimizes the average of the root mean square errors for the sample sizes considered, over a fine

| $(m,n)$ | $M_1$ | | $M_2$ | | $M_3$ | |
|---|---|---|---|---|---|---|
| | Method 1 | Method 2 ($0.04n^{-1/3}$) | Method 1 | Method 2 ($0.08n^{-1/5}$) | Method 1 | Method 2 ($0.04n^{-1/3}$) |
| (5 5) | 0.169, 0.045 | 0.180, 0.096 | 0.202, 0.116 | 0.218, 0.109 | 0.205, 0.075 | 0.237, 0.143 |
| (5 10) | 0.157, 0.067 | 0.166, 0.091 | 0.218, 0.172 | 0.213, 0.115 | 0.201, 0.109 | 0.208, 0.124 |
| (10 10) | 0.134, 0.033 | 0.141, 0.056 | 0.190, 0.139 | 0.193, 0.086 | 0.149, 0.046 | 0.156, 0.069 |
| (10 15) | 0.118, 0.049 | 0.126, 0.052 | 0.187, 0.155 | 0.190, 0.078 | 0.122, 0.053 | 0.129, 0.058 |
| (10 20) | 0.108, 0.062 | 0.109, 0.046 | 0.185, 0.167 | 0.176, 0.069 | 0.109, 0.064 | 0.110, 0.049 |
| (15 10) | 0.125, 0.018 | 0.126, 0.040 | 0.177, 0.117 | 0.184, 0.070 | 0.135, 0.020 | 0.132, 0.046 |
| (15 15) | 0.104, 0.038 | 0.109, 0.040 | 0.172, 0.140 | 0.175, 0.066 | 0.109, 0.038 | 0.112, 0.038 |
| (15 20) | 0.094, 0.042 | 0.098, 0.038 | 0.170, 0.149 | 0.174, 0.059 | 0.092, 0.044 | 0.100, 0.036 |
| (20 10) | 0.124, 0.010 | 0.123, 0.029 | 0.165, 0.112 | 0.175, 0.065 | 0.127, 0.007 | 0.123, 0.039 |
| (20 15) | 0.102, 0.025 | 0.106, 0.025 | 0.162, 0.133 | 0.170, 0.058 | 0.103, 0.026 | 0.106, 0.027 |
| (20 20) | 0.089, 0.033 | 0.097, 0.023 | 0.159, 0.139 | 0.161, 0.054 | 0.087, 0.036 | 0.093, 0.027 |
| (3 80) | 0.162, 0.145 | 0.105, 0.080 | 0.269, 0.262 | 0.164, 0.093 | 0.197, 0.166 | 0.110, 0.083 |
| (3 100) | 0.162, 0.146 | 0.099, 0.077 | 0.270, 0.265 | 0.159, 0.089 | 0.187, 0.159 | 0.098, 0.074 |
| (4 80) | 0.141, 0.124 | 0.094, 0.069 | 0.248, 0.242 | 0.157, 0.086 | 0.150, 0.129 | 0.098, 0.068 |
| (4 100) | 0.140, 0.125 | 0.088, 0.063 | 0.249, 0.244 | 0.148, 0.078 | 0.144, 0.125 | 0.087, 0.063 |

Table 2.1: Root mean square errors and biases , the first and second entries respectively, for the estimate of threshold $d_0$ obtained using Methods 1 and 2, for the three models with $\sigma = 0.3$ and different choices of $m$ and $n$.

grid.

The root mean square errors and the biases for each allocation pair are given in Table 2.1. Both procedures are inherently biased to the right as the $p$-values are not necessarily close to zero to the immediate right of $d_0$. When $m$ and $n$ are comparable, e.g., $m \leq 15$ and $n \leq 15$, Method 2, which relies on smoothing, does not perform well compared to Method 1. However, when $m$ is much smaller than $n$, e.g., $m = 4$ and $n = 80$, smoothing is efficient and Method 2 is preferred over Method 1. When both $m$ and $n$ are large, both methods work well. As Method 1 does not require selecting any tuning parameter, we recommend Method 1 in such situations.

### 2.1.6 Extension to Dependent Data

The global warming data falls under the standard regression setup, but involves dependent errors. Moreover, the data arises from a fixed design setting, with observations recorded annually. Here, we discuss the extension of Theorem 2.2 in this setting. With a fixed uniform design, we consider the model $Y_{i,n} = \mu\left(i/n\right) + \epsilon_{i,n}\ \ (i = 1, \ldots, n)$.

In such a model, $Y_{i,n}$ and $\epsilon_{i,n}$ must be viewed as triangular arrays. The estimator of the regression function is

$$\tilde{\mu}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \bar{Y}_{i\cdot} K\left(\frac{x - i/n}{h_n}\right).$$

For each $n$, we assume that the process $\epsilon_{i,n}$ is stationary and exhibits short-range dependence. Under Assumptions 1-5, listed in Robinson (1997), it can be shown that $\sqrt{nh_n}(\tilde{\mu}(x_k) - \mu(x_k))$, $x_k \in (0,1), k = 1,2$ and $x_1 \neq x_2$, converge jointly in distribution to independent normals with zero mean. In this setting, the working $p$-values, defined here to be $p_n^{(1)}(x, \tau_0) = 1 - \Phi(\sqrt{nh_n}(\tilde{\mu}(x) - \tau_0))$, still exhibit the desired dichotomous behavior. To keep the approach simple, we have not normalized by the estimate of the variance as this would have involved estimating the auto-correlation function. The conclusions of Theorem 2.2 can be shown to hold when $\hat{d}_n$ is constructed using (2.4) based on $p_n^{(1)}(X_i, \hat{\tau})$s. Here, $\hat{\tau}$ is constructed via averaging the responses over an interval that can be safely assumed to be on the left of $d_0$, as discussed in Section 2.1.4.

## 2.2 Simulation Results and Data Analysis

### 2.2.1 Simulation Studies

We consider the same three choices of the regression function $M_1$, $M_2$ and $M_3$, as in Section 2.1.5. The data are generated for allocation pair $(m, n)$ and a choice of regression function, with the errors being independent $N(0, \sigma^2)$, where $\sigma = 0.3$. The $X_i$s are again sampled from Uniform(0,1). We study the performance of the two methods when the estimates of $d_0$ are constructed using $p$-values that are normalized by their respective estimates of variances.

Firstly, we consider Method 1. In Table 2.2, we report the root mean square error and the bias for the estimators of $d_0$ and $\tau_0$, for different choices of $m$ and $n$. For moderate sample sizes, $M_3$ shows greater root mean square errors in general than $M_1$

| $(m,n)$ | $M_1$ | | $M_2$ | | $M_3$ | |
|---|---|---|---|---|---|---|
| | $d_0$ | $\tau_0$ | $d_0$ | $\tau_0$ | $d_0$ | $\tau_0$ |
| (5, 5) | 0.255, 0.215 | 0.175, 0.099 | 0.282, 0.255 | 0.134, 0.060 | 0.312, 0.262 | 0.142, 0.084 |
| (5, 10) | 0.248, 0.205 | 0.143, 0.086 | 0.271, 0.223 | 0.102, 0.049 | 0.303, 0.243 | 0.112, 0.072 |
| (10, 10) | 0.207, 0.157 | 0.124, 0.067 | 0.246, 0.216 | 0.077, 0.035 | 0.272, 0.215 | 0.104, 0.069 |
| (10, 20) | 0.172, 0.139 | 0.090, 0.052 | 0.240, 0.224 | 0.054, 0.029 | 0.248, 0.198 | 0.086, 0.062 |
| (10, 50) | 0.136, 0.121 | 0.056, 0.038 | 0.235, 0.228 | 0.038, 0.027 | 0.186, 0.157 | 0.070, 0.058 |
| (20, 50) | 0.090, 0.076 | 0.031, 0.018 | 0.194, 0.187 | 0.025, 0.017 | 0.124, 0.100 | 0.050, 0.034 |
| (50, 100) | 0.050, 0.043 | 0.011, 0.007 | 0.152, 0.148 | 0.012, 0.009 | 0.052, 0.046 | 0.014, 0.009 |

Table 2.2: Root mean square errors and biases , the first and second entries respectively, for the estimate of threshold $d_0$ obtained using Method 1 and the estimate of $\tau_0$ with $\sigma = 0.3$ for the three models.

and $M_2$ as the signal is weak close to 1 for $M_3$. For large sample sizes, the performance of the estimate is similar for $M_1$ and $M_3$ and is better than that for $M_2$, which can be ascribed to $M_2$ being smoother at $d_0$. The procedure is inherently biased to the right as $p$-values are not necessarily close to zero to the immediate right of $d_0$. Further, the estimator, on average, moves to the left with increase in $m$ as the desired dichotomous behavior becomes more prominent.

Next, we study the performance of Method 2. As the estimation procedure is entirely based on $\{(X_i, \bar{Y}_{i\cdot})\}_{i=1}^n$, without loss of generality, we take $m$ to be 1. We again work with the Gaussian kernel with the smoothing bandwidth chosen in the same fashion as in Section 2.1.5. In Table 2.3, we report the root mean square error and the bias for the two estimators, for different choices of $m$ and $n$. We see trends similar to those for Method 1, across the choices of the regression functions.

### 2.2.2 An allocation problem

In common dose-response studies, one is given a *total budget* of $N \equiv n \times m$ samples that need to be *allocated* to $n$ covariate values and $m$ replicates at each covariate value, respectively. Intuitively, increasing the number of replicates $m$ decreases the bias, whereas increasing the number of values $n$ of the covariate, decreases the variance of the estimators. The optimal allocation occurs when the two terms are balanced,

| $n$ | $M_1$ $h_n = 0.1n^{-1/3}$ | | $M_2$ $h_n = 0.15n^{-1/5}$ | | $M_3$ $h_n = 0.1n^{-1/3}$ | |
|---|---|---|---|---|---|---|
| | $d_0$ | $\tau_0$ | $d_0$ | $\tau_0$ | $d_0$ | $\tau_0$ |
| 20 | 0.285, 0.179 | 0.209, 0.105 | 0.290, 0.178 | 0.147, 0.057 | 0.326, 0.224 | 0.174, 0.084 |
| 30 | 0.268, 0.155 | 0.184, 0.094 | 0.268, 0.146 | 0.122, 0.038 | 0.319, 0.218 | 0.151, 0.074 |
| 50 | 0.237, 0.138 | 0.158, 0.080 | 0.244, 0.124 | 0.099, 0.031 | 0.284, 0.187 | 0.131, 0.069 |
| 80 | 0.215, 0.112 | 0.137, 0.066 | 0.222, 0.084 | 0.078, 0.019 | 0.270, 0.178 | 0.117, 0.068 |
| 100 | 0.195, 0.096 | 0.125, 0.053 | 0.216, 0.082 | 0.075, 0.017 | 0.251, 0.147 | 0.109, 0.061 |
| 200 | 0.159, 0.062 | 0.088, 0.035 | 0.191, 0.060 | 0.049, 0.011 | 0.210, 0.122 | 0.092, 0.053 |
| 500 | 0.104, 0.006 | 0.046, 0.014 | 0.164, 0.039 | 0.027, 0.005 | 0.142, 0.054 | 0.060, 0.025 |
| 1000 | 0.095, 0.004 | 0.031, 0.007 | 0.150, 0.020 | 0.020, 0.004 | 0.105, 0.021 | 0.039, 0.012 |
| 1500 | 0.085, 0.003 | 0.023, 0.005 | 0.148, 0.015 | 0.018, 0.003 | 0.088, 0.008 | 0.028, 0.008 |
| 2000 | 0.072, 0.002 | 0.020, 0.005 | 0.138, 0.007 | 0.015, 0.002 | 0.081, 0.001 | 0.023, 0.005 |

Table 2.3: Root mean square errors and biases , the first and second entries respectively, for the estimate of threshold $d_0$ obtained using Method 2 and the estimate of $\tau_0$ with $\sigma = 0.3$ for the three models.

| N | $M_1$ | | $M_3$ | |
|---|---|---|---|---|
| | $\sigma = 0.1$ | $\sigma = 0.3$ | $\sigma = 0.1$ | $\sigma = 0.3$ |
| 100 | (6,17) | (33,3) | (8,12) | (33,3) |
| 200 | (7,29) | (40,5) | (7,29) | (67,3) |

Table 2.4: Optimal allocation $(m, n)$ pairs for a fixed total budget $N = m \times n$

usually at a moderate value of $n$ and $m$, which depends on the value of $\sigma$ and the regression function. Thus, for a fixed $N$, one expects that the root mean square error exhibit a U-shape as a function of $m$; further, for larger $\sigma$ the optimal allocation would occur at a larger value of $m$.

We investigate this allocation problem for Method 1 through a simulation study and we present the optimal allocations for models $M_1$ and $M_3$. The setting under consideration is $d_0 = 0.5$, $N = 100$ and 200 and $\sigma = 0.1$ and 0.3. All possible combinations of $m$ and $n$ that approximately satisfy the total budget were considered. As very small values of $n$ are also considered, we work with a discrete uniform design for this study. The optimal allocations are shown in Table 2.4. It can be seen that for small $\sigma$, lots of covariate values and fewer replicates are preferred, while the situation gets reversed for high $\sigma$. Further, qualitatively similar results, in accordance with our observation above, are obtained for the other models. Nevertheless, a few anomalies are present; specifically, as we are sampling from the discrete uniform design on $[0, 1]$,

Table 2.5: Root mean square errors for the five procedures for different choices of $m$ and $n$ and models $M_1$ and $M_2$ when $\sigma = 0.3$.

| $(m, n)$ | $M_1$ | | | | | $M_2$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ |
| (5, 5) | 0.163 | 0.207 | 0.339 | 0.255 | 0.299 | 0.204 | 0.241 | 0.420 | 0.291 | 0.298 |
| (5, 10) | 0.134 | 0.176 | 0.304 | 0.307 | 0.344 | 0.201 | 0.227 | 0.390 | 0.346 | 0.360 |
| (10, 10) | 0.119 | 0.120 | 0.227 | 0.228 | 0.328 | 0.168 | 0.194 | 0.334 | 0.302 | 0.360 |
| (10, 20) | 0.092 | 0.079 | 0.191 | 0.265 | 0.295 | 0.177 | 0.163 | 0.303 | 0.329 | 0.354 |
| (10, 50) | 0.085 | 0.042 | 0.179 | 0.310 | 0.247 | 0.193 | 0.150 | 0.294 | 0.369 | 0.332 |
| (20, 50) | 0.060 | 0.030 | 0.128 | 0.212 | 0.176 | 0.162 | 0.147 | 0.245 | 0.305 | 0.274 |
| (50, 100) | 0.038 | 0.013 | 0.080 | 0.142 | 0.114 | 0.132 | 0.145 | 0.197 | 0.254 | 0.211 |

and $d_0 = 0.5$, sometimes the optimal allocation occurs at the rather extreme value $n = 3$. This is due to the fact that in that case, the covariate values are placed at 0.25, 0.5 and 0.75, and when $m$ is large, the fitted break point $\hat{d}_n$ is usually 0.5, the true parameter value. Whenever this is the case, the estimation error is exactly zero, making the observed root mean square error small. With the same budget, a larger $n$, say $n = 5$, can also lead to 0.5 as a covariate value, but the value of $m$ decreases in the process thereby increasing the bias and there are more options for the fitted break point to differ from 0.5, leading to larger root mean square errors.

### 2.2.3 Comparison with other procedures from dose-response setting

We now compare Method 1 to some competing procedures developed in the pharmacological dose-response setting to identify the minimum effective dose. Most of the methods developed in dose-response setting context are based on hypothesis testing procedures. For example, Williams (1971) developed a method to identify the lowest dose at which there is activity in toxicity studies using a closed testing procedure based on isotonic regression for a monotone dose-response relationship. Hsu and Berger (1999) developed a step-wise confidence set approach to estimate and make inference on the minimum effective dose. A nonparametric method based on the Mann–Whitney statistic incorporating the step-down procedure is investigated in Chen (1999), while Tamhane and Logan (2002) use multiple testing procedures for the task at hand. We compare our method with that of Williams (1971), of Hsu and Berger (1999) and of Chen (1999), referred henceforth as $P_3, P_4$ and $P_5$, respectively.

We compare the performance of Method 1, which we refer to as $P_1$, with that of $P_3$, $P_4$ and $P_5$. A natural parametric procedure to estimate $d_0$ might be to fit a kink-type model like $M_1$ to the observed responses and estimate $d_0$ and the slope of the linear segment by the least squares method. We also implement this method and call it $P_2$. Obviously when the true underlying regression function $\mu$ is not a kink-model this method might not be consistent, but given a finite sample it is often a good first approximation. Whereas, when $\mu$ is a kink-function, e.g., when we assume the true model to be $M_1$, this approach should clearly outperform the other procedures. Indeed, Table 5 shows that $P_2$ is very competitive for model $M_1$; still our approach $P_1$ performs better for small sample sizes, e.g., (5, 5) and (5, 10). For the model $M_2$, a slight departure from the model $M_1$, $P_1$ mostly dominates $P_2$, and all the other procedures. Note that as $P_3$, $P_4$ and $P_5$ are procedures that are based on testing hypotheses, we need to specify a level $\alpha$, and in the simulations reported in this chapter, we have set $\alpha = 0.05$. The choice of the $\alpha = 0.05$ is purely based on classical hypothesis testing considerations; a proper choice of the tuning parameter is not available. Also, to implement $P_3$–$P_5$, we computed the cut-off values necessary to carry out the hypothesis tests using simulation, as such tables are not available for the different choices of $m$ and $n$ considered in this chapter.

Overall, $P_1$ is very competitive, and the simplicity of our approach coupled with its adaptivity to different types of mis-specifications, makes it a very attractive choice. Indeed, one of the novelties of our approach lies in the fact that we treat the estimation of $d_0$ purely as an estimation problem and not a result of a series of hypotheses tests, thereby avoiding the need to specify $\alpha$.

### 2.2.4 Some practical recommendations

Based on our simulation study, the following practical recommendations are in order. In terms of optimal allocation under a fixed budget $N$, it is better for one to invest in an increased number of covariate values $n$, rather than replicates $m$. Further, when the sample size is reasonably large, the procedure that avoids estimating the variance function and works with non-normalized $p$-values, is competitive and is recommended in the regression settings with heteroscedastic errors and time-series.

### 2.2.5 Data Applications

The first data application deals with a dose-response experiment that studies the effect on cells from the IPC-81 leukemia rat cell line to treatment with 1-methyl-3-butylimidazolium tetrafluoroborate, at different doses measured in $\mu$M, micro mols per liter (Ranke et al., 2004). The substance treating the cells is an ionic liquid and the objective is to study its toxicity in a mammalian cell culture to assess environmental hazards. The question of interest here is at what dose level toxicity becomes lethal and cell cultures stop responding.

It can be seen from the physiological responses shown in the left panel of Fig. 2.1, that there is a decreasing trend followed by a flat stretch. Hence, it is reasonable to postulate a response function that stays above a baseline level $\tau_0$ until a transition point $d_0$ beyond which it stabilizes at its baseline level. We assume errors to be heteroscedastic, as the variability in the responses changes with level of dose, with more variation for moderate dose levels compared to extreme dose levels. This is the small $(m, n)$ case with $m$ and $n$ being comparable; in fact, $m = n = 9$. Hence we apply *Method 1* to this problem. The estimate of $\tau_0$ was constructed using the procedure based on $p$-values as described in Section 2.1.4. We get $\hat{\tau} = 0.0286$ with the corresponding $\hat{d} = 5.522 \log \mu M$, the third observation from right. We believe that this is an accurate estimate of $d_0$, since the cell-cultures exhibit high responses at earlier dose levels and no significant signal to the right of the computed $\hat{d}$.

The second example, as discussed in the introduction, involves measuring mercury concentration in the atmosphere through the LIDAR technique. There are 221 observations with the predictor variable range varying from 390 to 720. As supported by the middle panel of Fig. 2.1, the underlying response function is at its baseline level followed by a steep descent, with the point of change being of interest. There is evidence of heteroscedasticity and hence, we employ Method 2 without normalizing by the estimate of the variance. It is reasonable to assume here that till the range value 480 the function is at its baseline. The estimate of $\tau$ is obtained by taking the average of observations until range reaches 480, which gives $\hat{\tau} = -0.0523$. The estimates $\hat{d}$, computed for bandwidths varying from 5 to 30, show a fairly strong agreement as they lie between 534 and 547, with the estimates getting bigger for larger bandwidths. The cross-validated optimal bandwidth for regression is 14.96 for which the corresponding estimate of $d_0$ is 541.

The global warming data contains global temperature anomalies, measured in degree Celsius, for the years 1850 to 2009. These anomalies are temperature deviations measured with respect to the base period 1961–1990. The data are modeled as described in Section 2.1.6. As can be seen in the right panel of Fig. 2.1, the function stays at its baseline value for a while followed by a non-decreasing trend. The flat stretch at the beginning is also noted in Zhao and Woodroofe (2012) where isotonic estimation procedures are considered in settings with dependent data. The estimate of the baseline value, after averaging the anomalies up to the year 1875, is $\hat{\tau} = -0.3540$. With the dataset having 160 observations, estimates of the threshold were computed for bandwidths ranging from 5 to 30. The estimates varied over a fairly small time frame, 1916–1921. This is consistent with the observation on page 2 of Zhao and Woodroofe (2012) that global warming does not appear to have begun until 1915. The optimal bandwidth for regression obtained through cross-validation is 13.56, for which $\hat{d}$ is 1920.

### 2.2.6 Extensions

Here we discuss some of the possible extensions of our proposed procedure.

*Fixed design setting:* Although the results in this chapter have been proven assuming a random design, they can be easily extended to a fixed design setup. Consistency of the procedures will continue to hold. Some of the these extensions, particularly in the standard regression setting, are considered in great detail in Chapters 3 and 4.

*Unequal replicates:* We primarily dealt with the case of a balanced design with a fixed number of replicates $m$ for every dose level $X_i$. The case of varying number of replicates $m_i$ can be handled analogously. In the dose-response setting, Theorem 2.1 will continue to hold provided the minimum of the $m_i$s goes to infinity. In the standard regression setting, Theorem 2.2 can also be generalized to the situation with unequal number of replicates at different doses.

## 2.3 Concluding Discussion

While we have developed a novel methodology for threshold estimation and established consistency properties rigorously, a pertinent question that remains to be addressed is the construction of confidence intervals for $d_0$. A natural way to approach this problem is to consider the limit distribution of our estimators for the two settings and use the quantiles of the limit distribution to build asymptotically valid confidence intervals. This is addressed in Chapters 3 and 4.

In this chapter, we have restricted ourselves to a univariate regression setup. Our approach can potentially be generalized to identify the baseline region, the set on which the function stays at its minimum, in multi-dimensional covariate spaces. This is a special case of the *edge estimation* problem, a problem of considerable interest in statistics and engineering. The $p$-values, constructed analogously, will continue to exhibit a limiting dichotomous behavior which can be exploited to construct estimates of the baseline region. Procedures that look for a jump in the derivative of a certain order of $\mu$ (Müller, 1992; Raimondo, 1998) do not have natural extensions to such high dimensional settings. We address this problem in Chapter 5.

# CHAPTER 3

# Asymptotics for the dose–response setting

For the model $Y = \mu(X) + \epsilon$ with $\mu : [0,1] \mapsto R$ satisfying

$$\mu(x) \begin{cases} = \tau_0 & \text{for } x \leq d_0 \\ > \tau_0 & \text{for } x > d_0; \end{cases}, \tag{3.1}$$

we proposed novel and computationally simple procedures for estimating $d_0$ in both the standard regression and dose-response settings in Chapter 2. We established consistency under mild conditions and also studied the finite sample properties of the estimates. However, the problem of constructing CIs for $d_0$ was not addressed. In this chapter, we address this inference question in the dose-response setting by deriving the asymptotic distribution of the estimator $\hat{d}_{m,n}$ (see (2.3)) as $m, n$ grow to infinity, and demonstrating how to use the quantiles of this distribution to set the limits of the CI. It turns out that the asymptotic behavior of the estimators in the dose-response setting is *fundamentally different* from that in the standard regression setting which we address separately in Chapter 4. The estimates in the regression setting converge to minimizers of processes with differentiable sample paths that can be written as transforms of Gaussian processes while, as we will see below, those in the dose-response setting converge to the minimizers of piecewise-constant processes with jump discontinuities. Thus, many of the tools that play a crucial role in standard regression setting are inapplicable in the dose-response case (see Remark 4.8 in Chapter 4).

It should be noted that the problem of estimating $d_0$ in different models has received much attention in the statistics literature. If $\mu$ is assumed to have a jump discontinuity at $d_0$, then $d_0$ corresponds to a usual change-point for $\mu$. Such change-point models are very well understood; see e.g., Hinkley (1970), Korostelëv (1987), Dümbgen (1991), Müller (1992), Korostelëv and Tsybakov (1993a), Loader (1996), Müller and Song (1997) and the references therein. Our results, here, are developed for the harder problem that arises when $\mu$ is *continuous* at $d_0$. In particular, the smoother the regression function in a neighborhood of $d_0$, the greater the challenge in estimating $d_0$ precisely. We show that if $d_0$ is a cusp of $\mu$ of order $k$ (i.e., the first $k-1$ right derivatives of $\mu$ at $d_0$ equal 0 but the $k$-th does not, so that $d_0$ is a change-point in the $k$-th derivative) and $\mu$ is locally monotone in a neighborhood of $d_0$, then $\hat{d}_{m,n} - d_0$ is of order $N^{-1/(2k+1)}$, where $N = m \times n$ is the total budget and $m$ is chosen in some optimal manner (to be specified later) in terms of $n$.

The limit distribution of $N^{1/(2k+1)}(\hat{d}_{m,n} - d_0)$ is seen to be that of an appropriate minimizer of a jump process drifting off to infinity, that can be viewed as a generalization of a compound Poisson process. The derivation of the asymptotic distribution is complicated owing to the fact that the sample paths of the limit process are piecewise constant, resulting in non-unique minimizers. Hence, the more common continuous mapping arguments that rely on the uniqueness of the extremum of limit processes (see e.g., Theorem 2.7 of Kim and Pollard (1990)) – a phenomenon that shows up often with Gaussian limits and monotone transforms thereof – do not apply, and careful modifications, which rely on the continuity of the argmin functional in spaces of discontinuous functions, are required. In particular, the least squares estimate of $d_0$ (which is not unique) needs to be carefully picked. Another important challenge lies in deriving the rate of convergence of the estimator, which requires a considerable generalization of the standard rate theorems (see Theorem B.1) in the modern empirical processes literature (see e.g., Theorem 3.2.5 of van der Vaart and Wellner (1996)), and the choice of a cleverly constructed dichotomous metric on $\mathbb{R}$ (see Lemma B.2) to invoke the generalization. The details are available in the proof of Theorem 3.2.

The knowledge of $k$ is essential for constructing two-sided CIs based on these limiting results. Although resampling approaches such as subsampling are shown to work (in Section 3.3.2) for our problem, they do not present a solution for the situation when $k$ is unknown. We do end up providing a partial answer and show that adaptive upper confidence bounds can be constructed in the $k$-unknown case (Section 3.3.1).

The remainder of the chapter is organized as follows. In Section 3.1, we state the core assumptions and the variant of the $p$-based estimator that is primarily studied. The rates of convergence and the asymptotic distributions are deduced in Section 3.2, assuming a random design setting. Their implications to constructing CIs in practical applications, along with some auxiliary results on subsampling and adaptivity, are discussed in Section 3.3. In Section 3.4, we discuss the large sample behavior of the estimator of $d_0$ in a fixed design setting. We study the finite sample coverage performance of the CIs through simulations in Section 3.5 and discuss an application from a complex queuing system. Some conclusion are drawn in Section 3.6. The proofs of several technical results are provided in Appendix B.

## 3.1 Formulation and assumptions

For convenience, we study the problem in a random design setting with homoscedastic errors. The extension to the fixed design setting is considered in Section 3.4. The expression for the estimator of $d_0$ in fixed design setting is identical to that in the random design with the exception that the covariate $X_i$s would then just be fixed design points.

Assume the covariate $X$ is sampled from a Lebesgue density $f$, $X$ and $\epsilon$ are independent, $E(\epsilon) = 0$ and let $\sigma_0^2 := \text{Var}(\epsilon) > 0$. Consider data $\{(X_i, Y_{ij}) : 1 \leq j \leq m, 1 \leq i \leq n\}$, where the $X_i$s are i.i.d. random variables distributed like $X$, $\{\epsilon_{ij}\}$ are i.i.d. random variables distributed like $\epsilon$, the vectors $\{X_i\}$ and $\{\epsilon_{ij}\}$ are independent, and

$$Y_{ij} = \mu(X_i) + \epsilon_{ij}, \quad 1 \leq j \leq m, \text{ and } 1 \leq i \leq n. \tag{3.2}$$

Here, $N = m \times n$ is the total budget and we assume $m = m_0 n^\beta$ for some $\beta > 0$, to incorporate the scenario that $m$ can be 'large' relative to $n$, a feature of several dose-response studies.

Recall that $\bar{Y}_i = \sum_{i=1}^m Y_{ij}/m$, $\hat{\sigma}^2 = \sum_{i,j}(Y_{ij} - \bar{Y}_i)^2/(nm - n)$ and the normalized $p$-values are given by

$$p_{m,n}(X_i) = 1 - \Phi(\sqrt{m}(\bar{Y}_i - \tau_0)/\hat{\sigma}).$$

Let $\gamma = 3/4$ and $\mathbb{P}_n$ denote the empirical measure of $(X_i, \bar{Y}_i), i = 1, \ldots, n$. With a slight difference of notation from Chapter 2, let

$$\mathbb{M}_{m,n}(d) \equiv \mathbb{M}_{m,n}(d, \hat{\sigma}) = \mathbb{P}_n \left[ \left\{ \Phi\left( \frac{\sqrt{m}(\bar{Y} - \tau_0)}{\hat{\sigma}} \right) - \gamma \right\} 1(X \leq d) \right]. \tag{3.3}$$

As $(p_{m,n}(X_i) - 1/4) = (-1)\left(\Phi\left(\sqrt{m}(\bar{Y} - \tau_0)/\hat{\sigma}\right) - \gamma\right)$, the estimate from the criterion (2.3) of Chapter 2 is simply

$$\hat{d}_{m,n} = \operatorname*{sargmin}_{d \in [0,1]} \mathbb{M}_{m,n}(d).$$

Here, sargmin denotes the smallest argmin. We initially study this estimate assuming a known $\tau_0$. When $\tau_0$ is unknown, an estimate can be plugged in its place (more about this in Section 3.5). Also, for any choice of $\gamma \in (1/2, 1)$ in (3.3) the estimator of $d_0$ can be shown to be consistent by calculations similar to that in the proof of 2.1 from

Chapter 2.

The smoothness of the function in the vicinity of $d_0$ plays a crucial role in determining the rate of convergence. For the random design setting we make the following assumptions.

1. The regression function $\mu$ has a cusp of order $k$, $k$ being a *known* positive integer, at $d_0$, i.e., $\mu^{(l)}(d_0) = 0, 1 \leq l \leq k-1$ and $\mu^{(k)}(d_0+) > 0$, where $\mu^{(l)}(\cdot)$ denotes the $l$th derivative of $\mu$. Also, the $k$-th derivative, $\mu^{(k)}(x)$ is assumed to be continuous and bounded for $x \in (d_0, d_0 + \zeta_0]$ for some $\zeta_0 > 0$.

2. The errors $\epsilon$ possess a continuous positive density on a (finite or infinite) interval.

3. The design density $f$ for the dose-response setting is assumed to be continuous and positive on $[0, 1]$.

**Remark 3.1.** *Some words of explanation on why we address the asymptotics for a random design, as opposed to fixed design, are in order. It turns out that there is no limit distribution in this problem when the $X_i$s are the grid-points of a non-random grid, say, the uniform grid of size $n$, on the domain of the covariate. See Remark 3.7 for a more technical explanation of this issue. Moreover, note that our data application (see Section 3.5) does come from a random design.*

## 3.2 Main Results

We state and prove results on the limiting behavior of the estimator $\hat{d}_{m,n}$ discussed in Section 3.1. Results on the variants of the procedure discussed in Chapter 2 follow similarly and are stated without proofs in Section 3.2.3. The results in this section are developed for $\gamma \in (1/2, 1)$ and a known $\tau_0$. It will be seen in Section 3.3.3 that $\tau_0$ can be estimated at a sufficiently fast rate; consequently, even if $\tau_0$ is unknown, appropriate estimates can be substituted in its place to construct the $p$-values that are instrumental to the methods of this chapter, without changing the limit distributions. Without loss of generality, we take $\tau_0 \equiv 0$, as one can work with $(Y_{ij}-\tau_0)$s ($(Y_i-\tau_0)$s) in place of $Y_{ij}$s ($Y_i$s).

### 3.2.1 Rate of convergence

As $m = m_0 n^{\beta}$, we consider the asymptotics in the dose-response model as $n \to \infty$. Let $P_n$ denote the measure induced by $(\bar{Y}, X)$ and

$$M_{m,n}(d) = M_{m,n}(d,\sigma_0) = P_n\left[\left\{\Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma_0}\right) - \gamma\right\} 1(X \le d)\right].$$

The process $M_{m,n}$ is the population equivalent of $\mathbb{M}_{m,n}$ defined in (3.3) and can be simplified as follows. Let

$$Z_{1n} = \frac{1}{\sqrt{m}\sigma_0}\sum_{j=1}^{m} \epsilon_{1j} \tag{3.4}$$

and $Z_0$ be a standard normal random variable independent of $Z_{1n}$s. Then

$$\begin{aligned} E\left[\Phi\left(\frac{\sqrt{m}\bar{Y}_1}{\sigma_0}\right)\Bigg| X_1 = x\right] &= E\left[\Phi\left(\frac{\sqrt{m}\mu(x)}{\sigma_0} + Z_{1n}\right)\right] \\ &= E\left[E\left[1\left(Z_0 < \frac{\sqrt{m}\mu(x)}{\sigma_0} + Z_{1n}\right)\Bigg| Z_{1n}\right]\right] \\ &= P\left[\frac{Z_0 - Z_{1n}}{\sqrt{2}} < \frac{\sqrt{m}\mu(x)}{\sqrt{2}\sigma_0}\right] = \Phi_n\left(\frac{\sqrt{m}\mu(x)}{\sqrt{2}\sigma_0}\right), \end{aligned} \tag{3.5}$$

where $\Phi_n$ denotes the distribution function of $(Z_0 - Z_{1n})/\sqrt{2}$. Then, by integrating with respect to the density of $X$, it can be shown that

$$M_{m,n}(d) = \begin{cases} (\Phi_n(0) - \gamma)\, F(d), & d \le d_0, \\ (\Phi_n(0) - \gamma)\, F(d_0) + \int\limits_{d_0}^{d} \left[\Phi_n\left(\frac{\sqrt{m}\mu(x)}{\sqrt{2}\sigma_0}\right) - \gamma\right] f(x)dx, & d > d_0. \end{cases}$$

Let $d_{m,n} = \text{sargmin}_{d\in(0,1)} M_{m,n}(d)$. We first study the behavior of $d_{m,n}$ which satisfies

$$\Phi_n\left(\frac{\sqrt{m}\mu(d_{m,n})}{\sqrt{2}\sigma_0}\right) = \gamma.$$

Let $\Phi_n^{-1}$ be the left continuous inverse of $\Phi_n$. By Assumptions 1 and 2, we get

$$\frac{\mu^{(k)}(\zeta_n)}{k!}(d_{m,n} - d_0)^k = \frac{\sqrt{2}\sigma_0\Phi_n^{-1}(\gamma)}{\sqrt{m}}, \tag{3.6}$$

where $\zeta_n$ is some point between $d_0$ and $d_{m,n}$. As $n \to \infty$, the right-hand side (RHS) of the above display goes to zero. So, $d_{m,n} \to d_0$. Also, $\Phi_n$ converges point wise to $\Phi$ and the convergence holds for their inverse functions too. Hence,

$$d_{m,n} = d_0 + \left[\frac{k!\sqrt{2}\sigma_0\Phi^{-1}(\gamma)}{\mu^{(k)}(d_0+)}\right]^{1/k} m^{-1/(2k)} + o(m^{-1/(2k)}). \tag{3.7}$$

This shows that $d_{m,n} - d_0 = O(m^{-1/(2k)}) = O(n^{-\beta/(2k)})$. In a sense, $\hat{d}_{m,n}$, is estimating $d_{m,n}$ instead of $d_0$, and hence, its rate of convergence to $d_0$ can be expected to be at most of order $n^{-\beta/(2k)}$. Moreover, $\hat{d}_{m,n}$ is one of the order statistics of $X_i$s and hence, can only be close to $d_0$ up to an order $1/n$. We next provide a formal statement of the rate of convergence of $\hat{d}_{m,n}$.

**Theorem 3.2.** *Let $\alpha = min\,(1, \beta/(2k))$. Then,*

$$n^\alpha(\hat{d}_{m,n} - d_0) = m_0^{-\frac{\alpha}{1+\beta}} N^{\frac{\alpha}{1+\beta}}(\hat{d}_{m,n} - d_0) = O_P(1).$$

**Remark 3.3.** *The function $\mu$ may not satisfy Assumption 1 for any $k \in \mathbb{Z}$ and can still take off at $d_0$, e.g., $\mu_{(1)}(x) = \exp(-1/(x-d_0))1(x > d_0)$ and $\mu_{(2)}(x) = \exp(-1/(x-d_0)^2)1(x > d_0)$ are two such infinitely differentiable functions with a singularity at $d_0$. By calculations almost identical to those for deriving (3.7), it can be shown that $d_{m,n} - d_0 = O\left((\log(n))^{-1/i}\right)$ when $\mu = \mu_{(i)}$, $i = 1, 2$. Hence, we do not* *expect a universal rate of convergence for $\hat{d}_{m,n}$ when $\mu$ is infinitely differentiable at $d_0$ and adhere to Assumption 1.*

The proof is given in Section B.1 of Appendix B. The optimal rate corresponds to $\alpha = 1$. In terms of the total budget, the best possible rate is achieved when $\beta = 2k$. In that case, $N^{1/(2k+1)}(\hat{d}_{m,n} - d_0) = O_P(1)$. For, $\beta < 2k$, the rate of convergence is $n^{\beta/(2k)}$ or $N^{\beta/\{2k(1+\beta)\}}$.

**Remark 3.4.** *The rate $N^{-/1(2k+1)}$ is not surprising as it appears in inverse function estimation: for example, if $h$ is a smooth monotone function, the isotonic regression estimate of $x_0 := h^{-1}(\theta_0)$, where $\theta_0$ is a fixed point in the range of $h$, converges at rate $S^{-1/(2k+1)}$ ($S$ being sample size) under the assumption that $f$ is (at least) $k$-times differentiable at $x_0$, $f^{(k)}(x_0) \neq 0$ and $f^{(l)}(x_0) = 0$ for $1 \leq l < k$, which is the exact analogue of the 'cusp assumption' on $d_0$ above. We expect this rate to be minimax, even though a formal proof appears difficult and is outside the scope of this discussion; see Section 3.6 for more details.*

### 3.2.2 Asymptotic Distribution

We now deduce the asymptotic distribution of $\hat{d}_{m,n}$ for different choices of $\beta$, starting with $\beta = 2k$. Note that $n(\hat{d}_{m,n} - d_0) = \text{sargmin}_{t\in\mathbb{R}} \hat{V}_n(t)$ where

$$\hat{V}_n(t) = n\left\{\mathbb{M}_{m,n}\left(d_0 + t/n, \hat{\sigma}\right) - \mathbb{M}_{m,n}(d_0, \hat{\sigma})\right\}. \tag{3.8}$$

We deduce the limit of $\hat{V}_n$ and then apply a special continuous mapping theorem to obtain the asymptotic distribution of $\hat{d}_{m,n}$.

To state the limiting distribution, we introduce the following notation. Let $\{\nu^+(t) : t \geq 0\}$ and $\{\nu^-(t) : t \geq 0\}$ be two independent homogeneous Poisson processes with same intensity $f(d_0)$ but with RCLL (right continuous with left limits) and LCRL (left continuous with right limits) paths, respectively. Let $\{S_i\}_{i\geq 1}$ denote the arrival times for the process $\nu^+$. Further, let $\{Z_i\}_{i\geq 1}$ and $\{U_i\}_{i\geq 1}$ be independent sequences of i.i.d. $N(0,1)$'s and i.i.d. $U(0,1)$'s respectively which are, moreover, independent of the processes $\nu^+$ and $\nu^-$. Define $V(t)$ as:

$$V(t) = \begin{cases} \sum_{j=1}^{\nu^+(t)} \left( \Phi\left( \frac{\sqrt{m_0}\ \mu^{(k)}(d_0+)}{k!\sigma_0} S_j^k + Z_j \right) - \gamma \right), & t \geq 0, \\ \sum_{j=1}^{\nu^-(-t)} \left(\gamma - U_j\right), & t < 0, \end{cases} \tag{3.9}$$

where sum over a null set is taken to be zero. We will show that $\hat{V}_n$ converges weakly to $V$ as processes in $D(\mathbb{R})$, the space of càdlàg functions (right continuous having left limits) on $\mathbb{R}$ equipped with the Skorokhod topology; see Lindvall (1973) for more details on $D(\mathbb{R})$. Moreover, the asymptotic distribution of $\hat{d}_{m,n}$ will be characterized by a minimizer of the process $V$. The limiting process $V$ does not possess a unique minimizer as it stays at any level it attains for an exponential amount of time. Hence, the usual argmin (argmax) continuous mapping theorem (see for example Theorem 3.2.2 of van der Vaart and Wellner (1996)) does not suffice for deducing the limiting distribution; we also need to show the convergence of the involved jump processes (Lan et al., 2009, pp. 1760–1762).

For convenience, we state a consequence of Lemmas 3.1, 3.2 and 3.3 from Lan et al. (2009) which provides a version of the argmin (argmax) continuous mapping theorem required in our setting. Let $\mathcal{S}$ denote the class of piecewise constant functions in $D(\mathbb{R})$ that are continuous at every integer point, assume the value 0 at 0, and possess finitely many jumps on every compact interval $[-C, C]$, where $C > 0$ is an integer. Note that $\mathcal{S}$ is a closed subset of $D(\mathbb{R})$. Also, define the pure jump process, $\tilde{g}$, (of jump size 1) corresponding to the function $g \in D(\mathbb{R})$, as the piecewise constant right continuous function with left limits, such that for any $s > 0$, $\tilde{g}(s)$ counts the number of jumps of the function $g$ in the interval $[0, s]$, while for $s < 0$, $\tilde{g}(s)$ counts the number of jumps in the set $(s, 0)$. We have the following result.

**Theorem 3.5.** *Let $\mathbb{V}_n, n \geq 0$, be processes in $D(\mathbb{R})$ such that $\mathbb{V}_n \in \mathcal{S}$, with probability 1. Also, let $\mathbb{J}_n, n \geq 0$, denote the corresponding jump processes and $(\xi_n^s, \xi_n^l), n \geq 0$, be the smallest and largest minimizers for $\mathbb{V}_n$. Suppose that:*

*(i)* $(\mathbb{V}_n, \mathbb{J}_n)$ *converges weakly to* $(\mathbb{V}_0, \mathbb{J}_0)$ *as processes in* $D[-C, C] \times D[-C, C]$, *for each positive integer* $C$.

*(ii)* *No two flat stretches of* $\mathbb{V}_0(t), t \in [-C, C]$, *have the same height a.s., for each positive integer* $C$.

*(iii)* $\left\{(\xi_n^s, \xi_n^l), n \geq 0\right\}$ *is* $O_P(1)$.

*Then* $(\xi_n^s, \xi_n^l) \xrightarrow{d} (\xi_0^s, \xi_0^l)$, *where* $\xrightarrow{d}$ *denotes convergence in distribution.*

Note that $\hat{V}_n \in \mathcal{S}$ with probability 1. For $t \in \mathbb{R}$, let the function $\operatorname{sgn}(t)$ denote the sign of $t$. Also, let $J_n$ denote the jump process corresponding to $\hat{V}_n(t)$. Then,

$$J_n(t) = \operatorname{sgn}(t) \sum_{i=1}^{n} \left[ 1 \left( X_i \leq d_0 + \frac{t}{n} \right) - 1 \left( X_i \leq d_0 \right) \right].$$

Further, let $J$ be the jump process associated with $V(t)$, i.e., $J(t) = \nu^+(t) 1(t \geq 0) + \nu^-(-t) 1(t < 0)$. We have the following result.

**Theorem 3.6.** *Let $\beta = 2k$ and $\hat{V}_n$ and $V$ be as defined in* (3.8) *and* (3.9) *respectively. Then, the conditions (i), (ii) and (iii) of Theorem 3.5 are satisfied for $\mathbb{V}_n = \hat{V}_n$ and $\mathbb{V}_0 = V$ with $J_n$ and $J$ being the corresponding jump processes. As a consequence,*

$$n(\hat{d}_{m,n} - d_0) \xrightarrow{d} \operatorname*{sargmin}_{t \in \mathbb{R}} V(t).$$

The proof involves establishing finite dimensional convergence using characteristic functions and justifying a moment condition (see Billingsley (1968, pp. 128)) to prove asymptotic tightness. It is available in Section B.2.

**Remark 3.7.** *The counts* $\sum_{i \leq n} 1(X_i \in (d_0, d_0 + t/n])$ *account for the Poisson process that arises in the limit. If the* $X_i$*s were drawn from a fixed uniform design, these counts would not converge. Hence, a fixed design setup does not yield a limiting distribution for the underlying processes, and consequently for* $\hat{d}_{m,n}$*, in the dose-response setting. This fact was also observed in the change point setting of Lan et al. (2009, pp. 1766).*

The limiting random variable $\operatorname{sargmin}_{t \in \mathbb{R}} V(t)$ is continuous by virtue of the fact that the probability of a jump at a particular point for a Poisson process is zero. Its distribution depends upon the parameters $m_0$, $\mu^{(k)}(d_0+)$, $\sigma_0$, $f(d_0)$ and $\gamma$. It is clear from the expression for $V$ (see (3.9)) that a larger $m_0$, a larger $\mu^{(k)}(d_0+)$ or a smaller $\sigma_0$ will skew the limiting distribution more to the left. For the sake of completeness, we state the asymptotics for other choices of $\beta$. When $\beta > 2k$, the derivation of the limiting distribution is similar to that of Theorem 3.6 and is outlined in Section B.3 of Appendix B.

**Proposition 3.8.** *Let* $\beta > 2k$*. Also, let* $\{\nu_1^+(t) : t \geq 0\}$ *and* $\{\nu_1^-(t) : t \geq 0\}$ *be two independent homogeneous Poisson processes with same intensity* $f(d_0)$ *but with RCLL and LCRL paths respectively. Let* $\{\bar{U}_i\}_{i \geq 1}$ *be a sequence of i.i.d.* $U(0,1)$*s which is independent of* $\{\nu_1^+, \nu_1^-\}$*. Define* $\bar{V}(t)$ *as:*

$$\bar{V}(t) = \begin{cases} (1-\gamma)\nu_1^+(t), & t \geq 0, \\ \displaystyle\sum_{j=1}^{\nu_1^-(-t)} \left(\gamma - \bar{U}_j\right), & t < 0, \end{cases}$$

*where sum over a null set is taken to be zero. Then,* $n(\hat{d}_{m,n} - d_0) \xrightarrow{d} \operatorname{sargmin}_{t \in \mathbb{R}} \bar{V}(t) = \operatorname{sargmin}_{t \leq 0} \bar{V}(t)$.

The case $\beta < 2k$ yields a markedly different result from the above two scenarios: we do not get a non-degenerate limiting distribution any longer as the normalized estimator converges to a constant. The proof is given in Section B.4 of Appendix B.

**Proposition 3.9.** *Choose $\beta < 2k$. Let*

$$\hat{H}_n(t) = n^{\beta/(2k)} \left\{ \mathbb{M}_{m,n} \left( d_0 + \frac{t}{n^{\beta/(2k)}}, \hat{\sigma} \right) - \mathbb{M}_{m,n}(d_0, \hat{\sigma}) \right\}$$

*and*

$$c(t) = \begin{cases} \left(\frac{1}{2} - \gamma\right) f(d_0) t, & t \le 0 \\ f(d_0) \int\limits_0^t \left\{ \Phi \left( \frac{\sqrt{m_0}\ \mu^{(k)}(d_0+)}{\sqrt{2} k! \sigma_0} u^k \right) - \gamma \right\} du, & t > 0. \end{cases}$$

*Then, for any $L > 0$,*

$$\sup_{t \in [-L, L]} |\hat{H}_n(t) - c(t)| \xrightarrow{P} 0, \tag{3.10}$$

*and*

$$n^{\beta/(2k)}(\hat{d}_{m,n} - d_0) \xrightarrow{P} \operatorname*{argmin}_{t \in \mathbb{R}} \{c(t)\} = \left( \frac{\sqrt{2}\ k!\ \sigma_0\ \Phi^{-1}(\gamma)}{\sqrt{m_0}\ \mu^{(k)}(d_0+)} \right)^{1/k}.$$

### 3.2.3 Limit distributions for variants of the procedure

The rates of convergence and asymptotic distributions can be obtained similarly for the variants of the procedure that were discussed in Chapter 2. In what follows, we state the limiting distributions, without proofs, for one of the variants that was studied in detail in Chapter 2.

For heteroscedastic errors, the non-normalized version of the procedure ($p$-values are not normalized by the estimate of the variance), yields the following limiting distribution.

**Proposition 3.10.** *Consider the dose-response setting with heteroscedastic errors, i.e., $\sigma_0^2(x) = Var(\epsilon \mid X = x)$ need not be identically $\sigma_0$ but is assumed to be continuous and positive. Let*

$$\tilde{d}_{m,n} = \operatorname*{sargmin}_{d \in (0,1)} \mathbb{P}_n \left[ \left\{ \Phi\left( \sqrt{m}(\bar{Y} - \tau_0) \right) - \gamma \right\} 1(X \le d) \right],$$

*with* $m = m_0 n^{2k}$. *Let* $\{\nu^+(t) : t \geq 0\}$ *and* $\{\nu^-(t) : t \geq 0\}$ *be two independent homogeneous Poisson processes with same intensity* $f(d_0)$ *but with RCLL and LCRL paths, respectively. Let* $\{S_i\}_{i\geq 1}$ *denote the arrival times for the process* $\nu^+$. *Further, let* $\left\{Z_i^{(1)}\right\}_{i\geq 1}$ *and* $\left\{Z_i^{(2)}\right\}_{i\geq 1}$ *be independent sequences of i.i.d.* $N(0, \sigma_0^2(d_0))$*'s. Define* $\tilde{V}(t)$ *as:*

$$\tilde{V}(t) = \begin{cases} \displaystyle\sum_{j=1}^{\nu^+(t)} \left\{ \Phi\left( \frac{\sqrt{m_0}\ \mu^{(k)}(d_0+)}{k!} S_j^k + Z_j^{(1)} \right) - \gamma \right\}, & t \geq 0, \\ \displaystyle\sum_{j=1}^{\nu^-(-t)} \left\{ \gamma - \Phi(Z_j^{(2)}) \right\}, & t < 0, \end{cases}$$

*Then,* $n(\tilde{d}_{m,n} - d_0) \xrightarrow{d} \operatorname{sargmin}_{t\in\mathbb{R}} \tilde{V}(t)$.

## 3.3 Construction of CIs

As the form of the limit distribution depends upon the allocation of the total budget $N$ between $m$ and $n$ and may involve $k$, the construction of CIs requires some care. Consider, first, the case that $k$ is assumed known. Writing $m = m_0\, n^\beta$, we can set $\beta = 2k$, the optimal choice in terms of the total budget, to solve for $m_0$, and then construct a CI for $d_0$ using the result in Theorem 3.6. This requires estimating nuisance parameters like $f(d_0)$, $\sigma_0$ and $\mu^{(k)}(d_0+)$, of which the last is the hardest to estimate. Note that we have already estimated $\sigma_0$ in order to construct $\hat{d}_{m,n}$, while the design density at $d_0$ can be estimated using $\hat{f}(\hat{d}_{m,n})$, where $\hat{f}$ is a standard kernel density estimate of $f$. As far as $\mu^{(k)}(d_0+)$ is concerned, observe that

$$\mu(x) = \mu^{(k)}(d_0+)(x - d_0)^k/k! + o((x - d_0)^k)$$

for $x > d_0$. An estimate of $\mu^{(k)}(d_0+)$ can, therefore, be obtained by fitting a local polynomial to the right of $\hat{d}_{m,n}$ that involves the $k$-th power of the covariate.

Specifically, an estimate of $\xi_0 \equiv \mu^{(k)}(d_0+)/k!$ is:

$$\begin{aligned}\hat{\xi} &= \operatorname*{argmin}_{\xi} \sum_{i=1}^{n} \{\bar{Y}_i - \xi(X_i - \hat{d}_{m,n})^k\}^2 1(X_i \in (\hat{d}_{m,n}, \hat{d}_{m,n} + b_n]) \\ &= \frac{\sum \bar{Y}_i (X_i - \hat{d}_{m,n})^k 1(X_i \in (\hat{d}_{m,n}, \hat{d}_{m,n} + b_n])}{\sum (X_i - \hat{d}_{m,n})^{2k} 1(X_i \in (\hat{d}_{m,n}, \hat{d}_{m,n} + b_n])},\end{aligned}$$

where $b_n \downarrow 0$ and $nb_n^{2k+1} \to \infty$. The condition $nb_n^{2k+1} \to \infty$ is typical for estimating the $k$-th derivative at a known fixed point; see e.g., Gasser and Müller (1984), Härdle and Gasser (1985). The following lemma, whose proof is given in Section B.5 of Appendix B, justifies the consistency of this estimate for the optimal choice of $\beta$, thereby providing a way to construct CIs by imputing this estimate in the limiting distribution.

**Proposition 3.11.** *Let $\beta = 2k$. Then $\hat{\xi} \xrightarrow{P} \xi_0$.*

**Remark 3.12.** *The estimate $\hat{\xi}$ is effectively a kernel estimate with the smoothing kernel being uniform on $(0, 1]$. Alternative consistent estimators of $\xi$ can be obtained using other one sided kernels. To fix ideas, we only use the above mentioned estimate in the chapter.*

### 3.3.1 Adaptive upper confidence bounds

Note that the above inference strategy is not adaptive to the order of smoothness, $k$, at $d_0$. While we have not been able to develop an adaptive method for two-sided CIs, we are able to propose a strategy for one-sided honest CIs for $d_0$ (which are also of consequence in applications) that avoids knowledge of $k$. For example, if $d_0$ represents the minimal effective dose in a pharmacological setting, practitioners would be naturally interested in finding an upper confidence bound for $d_0$. The following result, whose proof follows along the same lines as that of Proposition 3.8, is our starting point for building such CIs.

**Proposition 3.13.** *Consider the dose-response setting with homoscedastic errors and normalized p–values and define* $\underline{d}_{m,n} = \mathrm{sargmin}_{0\le d\le d_0} \mathbb{M}_{m,n}(d)$*, so that* $\underline{d}_{m,n} \le \hat{d}_{m,n} = \mathrm{sargmin}_{0\le d\le 1} \mathbb{M}_{m,n}(d)$*. Then, for any* $\beta > 0$*,*

$$n\left(\underline{d}_{m,n} - d_0\right) \Rightarrow_d \underset{t\le 0}{\mathrm{sargmin}} \sum_{j=1}^{\nu^-(t)} \left(\gamma - U_j\right),$$

*where* $U_j$*'s and* $\nu^-$ *are as in Theorem 3.6.*

In fact the above result does not require $m$ to grow as a power of $n$. The condition $\min(m,n) \to \infty$ suffices. Note that the limit distribution above is concentrated on the negative axis (as it must, since $\underline{d}_{m,n} \le d_0$) and does not depend upon $k$. Simulating its quantiles requires just an estimate of $f(d_0)$. Let $K_\alpha$ be its $\alpha$'th quantile. Then,

$$\lim_{n\to\infty} P\left(d_0 \le \underline{d}_{m,n} - K_\alpha/n\right) = 1 - \alpha\,.$$

Now, $\underline{d}_{m,n}$ is obviously unknown, but $\underline{d}_{m,n} \le \hat{d}_{m,n}$ which is known. It follows easily that:

$$\liminf_{n\to\infty} P\left(d_0 \le \hat{d}_{m,n} - K_\alpha/n\right) \ge 1 - \alpha\,.$$

An essentially honest level $1-\alpha$ upper confidence bound for $d_0$ is therefore given by $[0, \hat{d}_{m,n} - K_\alpha/n]$. For an asymptotic allocation where $\beta > 2k$, by Proposition 3.8, the limit distributions of $\underline{d}_{m,n}$ and $\hat{d}_{m,n}$ coincide. Hence, these conservative upper confidence bounds are in a sense, minimally conservative, as they are exact for the situation $\beta > 2k$.

### 3.3.2 Subsampling

As an alternative to using the limit distribution, subsampling can be used to construct CIs for the case $\beta \ge 2k$. Let $q_n$ be a sequence of integers such that $q_n/n \to 0$ and $q_n \to \infty$. A subsample is constructed by selecting $q_n$ many $X_i$s and $l_n = \lfloor q_n m/n \rfloor$

response values at each selected $X_i$. The subsamples are denoted by $\mathcal{S}_1, \ldots, \mathcal{S}_{N_n}$, where $N_n = \binom{n}{q_n}\left[\binom{m}{l_n}\right]^{q_n}$. Let $\hat{d}_{n,q_n,j}$ denote the estimate of $d_0$ based on $\mathcal{S}_j, j = 1, \ldots, N_n$. Let $G_{n,\beta}$ denote the distribution of $n(\hat{d}_{m,n} - d_0)$. For $\beta \geq 2k$, $G_{n,\beta}$ converges weakly to a continuous limiting distribution, say $G_\beta$. The approximation to $G_{n,\beta}$, based on subsampling, is given by

$$L_{n,q}(x) = L_{n,q}(x, \beta) = \frac{1}{N_n} \sum_{j=1}^{N_n} 1\left[q_n(\hat{d}_{n,q_n,j} - \hat{d}_{m,n}) \leq x\right].$$

The following result justifies the use of subsampling in constructing CIs for $d_0$.

**Proposition 3.14.** *Let $\beta \geq 2k$. If $q_n/n \to 0$ and $q_n \to \infty$ then:*

*(i)* $\sup_x |L_{n,q}(x, \beta) - G_\beta(x)| \xrightarrow{P} 0$.

*(ii)* $P[c_{n,q,\alpha/2} \leq n(\hat{d}_{m,n} - d_0) \leq c_{n,q,1-\alpha/2}] \to 1 - \alpha$, *where* $c_{n,q,\xi} = \inf\{x : L_{n,q}(x) \geq \xi\}$.

The proof follows along the lines of that of Theorem 15.7.1 in Lehmann and Romano (2005). The details are provided in Section B.6 of Appendix B. The usual bootstrap methodology is not expected to be consistent.

### 3.3.3 The case of an unknown $\tau_0$

While our results have been deduced under the assumption of a known $\tau_0$, in real applications $\tau_0$ is generally not known. In this situation, quite a few extensions are possible. If $d_0$ can be safely assumed to be larger than some $\eta$, then a simple averaging of the observations below $\eta$ would yield a $\sqrt{mn}$-consistent estimator of $\tau_0$. If a proper choice of $\eta$ is not available, one can obtain an initial estimate of $\tau_0$ using the method proposed in Section 2.1.4 of Chapter 2, compute $\hat{d}_{m,n}$ and then average the responses from, say, $[0, c\hat{d}_{m,n}]$, $c \in (0, 1)$, to obtain an estimate of $\tau_0$, which will also be $\sqrt{mn}$-consistent. Note that this leads to an iterative procedure which we discuss in more detail in Section 3.5.1. Using a $\sqrt{mn}$-consistent estimate of $\tau_0$, say $\hat{\tau}$, so that the $\overline{Y}_i$s are centered around $\hat{\tau}$ in the $p$-values, it can be shown that all the asymptotic results encountered earlier stay unchanged. A brief sketch of the following result is given in Section B.7.

**Proposition 3.15.** *Let $\hat{d}_{m,n}$ now denote the smallest minimizer of*

$$\mathbb{M}_{m,n}(d, \hat{\sigma}, \hat{\tau}) = \mathbb{P}_n \left[ \left\{ \Phi\left( \frac{\sqrt{m}(\bar{Y} - \hat{\tau})}{\hat{\sigma}} \right) - \gamma \right\} 1(X \leq d) \right],$$

*where $\sqrt{mn}\,(\hat{\tau} - \tau_0) = O_p(1)$. For $m = m_0 n^{\beta}$ and $\alpha$ as defined in Theorem 3.2, we have $n^{\alpha}(\hat{d}_{m,n} - d_0) = O_P(1)$. Also, when $\beta = 2k$,*

$$n(\hat{d}_{m,n} - d_0) \xrightarrow{d} \operatorname*{sargmin}_{t \in \mathbb{R}} V(t),$$

*where the process $V$ is as defined in* (3.9).

A similar extension of Proposition 3.13 is valid as well.

## 3.4 Fixed design setting

As mentioned in Section 3.1, the estimation procedure does not change when we move over from the random design setting to the fixed design setting. For example, with a model of the form

$$Y_{ij} = \mu\left(\frac{i}{n}\right) + \epsilon_{ij}, \quad 1 \leq i \leq n,\ 1 \leq j \leq m,$$

where $\epsilon_{ij}$s are independent and identically distributed with mean 0 and variance $\sigma_0^2$, an estimate for $d_0$, based on non-normalized $p$-values, is given by

$$\hat{d}^{FD}_{m,n} = \operatorname*{sargmin}_{d \in (0,1)} \mathbb{M}_{m,n}(d),$$

where

$$\mathbb{M}_{m,n}^{FD}(d) = \frac{1}{n}\sum_{i=1}^{n}\left\{\Phi\left(\sqrt{m}\bar{Y}_i\right) - \gamma\right\} 1\left(\frac{i}{n} \leq d\right).$$

Here, $\tau_0$ is assumed known and taken to be zero without any loss of generality. The following result, whose proof is outlined in Section B.8 of Appendix B, shows that $\hat{d}_{m,n}^{FD}$ attains the same rate of convergence as its counterpart in the random design setting.

**Proposition 3.16.** *For $m = m_0 n^\beta$ and $\alpha$ as defined in Theorem 3.2, we have*

$$n^\alpha(\hat{d}_{m,n}^{FD} - d_0) = O_P(1).$$

As mentioned in Remark 3.7, there is no limit distribution available in this setting as the sums of the form $\sum_i [1\,(i/n \leq d_0 + t/n) - 1\,(i/n \leq d_0)]$, $t \in \mathbb{R}$, do not converge. However, the asymptotic distributions obtained in the random design setup can be used for setting approximate CIs for $d_0$ in such cases. Section 5.1.2 of Lan et al. (2007) investigated this issue through simulations in the related setting of a change-point regression model where the quantiles of the limit distribution (of the least squares estimate of the change-point) in a uniform *random design* setting were used for constructing CIs for the change-point when the data were generated from a uniform *fixed design* setting. The CIs obtained were seen to have comparable lengths to those for data generated from the random setting but were prone to over-coverage, and were therefore honest in the fixed design setting. A similar phenomenon was observed in our problem.

## 3.5 Data analysis

### 3.5.1 Simulations

We consider the underlying regression function as $\mu(x) = [2(x - 0.5)]1(x > 0.5), x \in [0, 1]$, the curve $M_1$ from Chapter 2. This function is at its baseline value 0 up to $d_0 = 0.5$ and then rises to 1. The errors are assumed to be normally distributed with mean 0 and standard deviation $\sigma_0 = 0.1$. We work with $\gamma = 3/4$ as extreme values of $\gamma$ (close to 0.5 or 1) tend to cause instabilities. We study the coverage performance of the approximate CIs obtained from the limiting distributions with the nuisance parameters estimated.

We generate samples for different choices of $m$ and $n$, under $\mu$. The covariate $X$ is sampled from $U(0, 1)$. For estimation, the factor $m_0$ is chosen so that the allocation between $m$ and $n$ is optimum. We assume $\tau_0$ to be unknown and get its initial estimate through the $p$-value based approach proposed in Chapter 2 (see (2.5)). An iterative scheme is then implemented where we use this initial estimator of $\tau_0$ to compute $\hat{d}_{m,n}$, re-estimate $\tau_0$ by averaging the responses for which $X$ lies in $[0, 0.9\hat{d}_{m,n}]$ and proceed thus. On average, the estimates stabilize within 5 iterations. Firstly, we compare the distribution of $n(\hat{d}_{m,n} - d_0)$ for $m = n = 500$ data points over 5000 replications with the deduced asymptotic distribution. The Q-Q plot, shown in the left panel of Figure 3.5.1, reveals considerable agreement between the two distributions. In Table 3.1 we provide the estimated coverage probabilities of the CIs over 5000 replications for the model $\mu$ constructed by imputing estimates of the nuisance parameters (as discussed in Section 3.3) in the limiting distribution. The limiting process $V$ was generated over a compact set incorporating the fact that $d_0 \in (0, 1)$ and consequently $n(\hat{d}_{m,n} - d_0) \in [n(\hat{d}_{m,n} - 1), n\hat{d}_{m,n}]$. The smoothing bandwidth for estimating $\mu^{(k)}(d_0+)$ was chosen to be $5(n/\log n)^{-1/(2k+1)}$. The coverage performance is not very sensitive to the choice of this bandwidth as long as it is reasonably wide. The approximate CIs exhibit

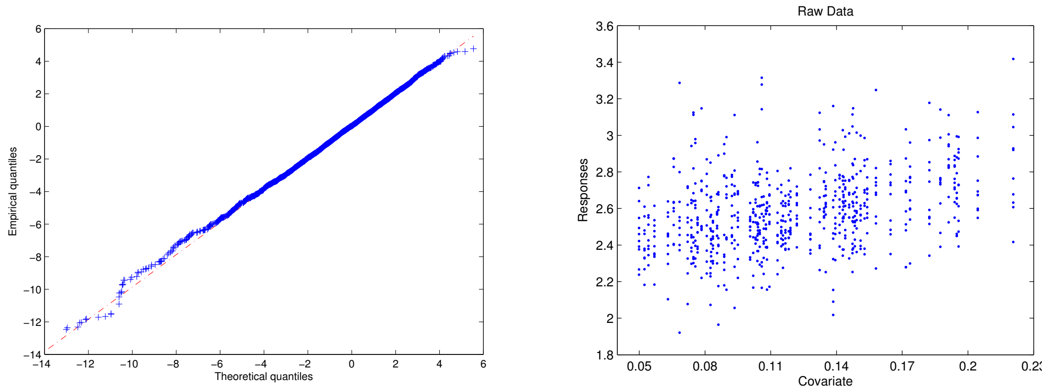



Figure 3.1: Q-Q plot under $\mu$ when $m = n = 500$ over 5000 replications (left plot), and the plot of all the response for the data from the queuing system (right plot).

over-coverage for small samples but have close to the desired nominal coverage level as the sample size increases. As discussed in Section 3.3.1, upper confidence bounds

| $m$ | $n$ | 90% CI | | 95% CI | |
|---|---|---|---|---|---|
| | | T | E | T | E |
| 5 | 5 | 0.966 (0.704) | 0.860 (0.637) | 0.973 (0.764) | 0.940 (0.696) |
| 10 | 10 | 0.941 (0.454) | 0.944 (0.473) | 0.970 (0.553) | 0.970 (0.568) |
| 15 | 10 | 0.924 (0.451) | 0.939 (0.472) | 0.966 (0.552) | 0.966 (0.564) |
| 10 | 15 | 0.914 (0.322) | 0.935 (0.338) | 0.961 (0.408) | 0.961 (0.428) |
| 15 | 15 | 0.913 (0.320) | 0.931 (0.345) | 0.959 (0.406) | 0.961 (0.435) |
| 20 | 20 | 0.910 (0.243) | 0.913 (0.254) | 0.955 (0.312) | 0.960 (0.326) |
| 25 | 25 | 0.908 (0.195) | 0.910 (0.202) | 0.951 (0.252) | 0.959 (0.259) |
| 30 | 30 | 0.903 (0.163) | 0.893 (0.167) | 0.951 (0.211) | 0.953 (0.215) |
| 50 | 50 | 0.901 (0.100) | 0.900 (0.100) | 0.950 (0.128) | 0.951 (0.130) |

Table 3.1: Coverage probabilities and lengths of two-sided CIs (in parentheses) using the true parameters (T) and the estimated parameters (E) for different sample sizes.

can be constructed without the knowledge of $k$. We provide coverage probabilities and average lengths of the CIs $[0, \hat{d}_{m,n} - K_{\alpha/n}]$, for $\alpha = 0.05$ and 0.10 in Table 3.2. The only parameter to estimate for computing the quantile $K_\alpha$ is $f(d_0)$ which, as mentioned earlier, is computed by evaluating a kernel estimate of $f$ at the point $\hat{d}_{m,n}$. As expected, the CIs are conservative but are close to the desired confidence level for large $m$ and $n$, with their average length converging towards 0.5 (length of the interval $[0, d_0]$).

| $m$ | $n$ | 90% CI | | 95% CI | |
|---|---|---|---|---|---|
| | | T | E | T | E |
| 5 | 5 | 0.951 (0.834) | 0.956 (0.865) | 0.970 (0.927) | 0.971 (0.930) |
| 10 | 10 | 0.955 (0.747) | 0.978 (0.753) | 0.990 (0.851) | 0.993 (0.857) |
| 15 | 10 | 0.962 (0.747) | 0.978 (0.750) | 0.990 (0.849) | 0.992 (0.855) |
| 10 | 15 | 0.933 (0.665) | 0.955 (0.672) | 0.972 (0.748) | 0.991 (0.754) |
| 15 | 15 | 0.921 (0.657) | 0.959 (0.669) | 0.966 (0.741) | 0.990 (0.751) |
| 20 | 20 | 0.920 (0.618) | 0.943 (0.627) | 0.962 (0.680) | 0.986 (0.690) |
| 25 | 25 | 0.921 (0.594) | 0.934 (0.598) | 0.960 (0.644) | 0.972 (0.649) |
| 30 | 30 | 0.915 (0.579) | 0.935 (0.584) | 0.960 (0.620) | 0.971 (0.626) |
| 50 | 50 | 0.913 (0.548) | 0.933 (0.551) | 0.958 (0.573) | 0.970 (0.576) |

Table 3.2: Coverage probabilities and lengths of one-sided adaptive CIs (in parentheses) using the true parameters (T) and the estimated parameters (E) for different sample sizes.

### 3.5.2 Complex queuing system

We consider a complex queuing system comprising multiple classes of customers waiting at infinite capacity queues and a set of processing resources modulated by an external stochastic process. This data is preferred over the toxicology data in Chapter 2 from the perspective of large sample inference. The system employs a resource allocation (scheduling) policy that decides at every time slot which customer class to serve, given the state of the modulating rate process and the backlog of the various queues. In Bambos and Michailidis (2004), a low complexity policy was introduced and its maximum throughput properties established. This canonical system captures the essential features of data/voice transmissions in a wireless network, in multi-product manufacturing systems, and in call centers (for more details see Bambos and Michailidis (2004)). An important quantity of interest to the system's operator is the average delay of jobs (over all classes), which constitutes a key performance metric of the quality of service offered by the system. The average delay of the jobs in a two-class system as a function of its loading under the optimal policy, for a small set of loadings is shown in the right panel of Figure 3.5.1. These responses were obtained through simulation, since for such complex systems analytic calculations of delays are intractable. More specifically, ten replicates of the response (average delay) were obtained based on 5,000 events per class by simulating the system under consideration and after accounting for a burn-in period of 2,000 per class in order to ensure that it reached its stationary regime. The means per loading, $\bar{Y}_i$s, are shown in the left panel of Figure 3.5.2. The system operator is interested in identifying the loading beyond which the average delay starts increasing from its initial baseline value. Starting with an initial estimate of $\tau_0$, the iterative approach discussed in the

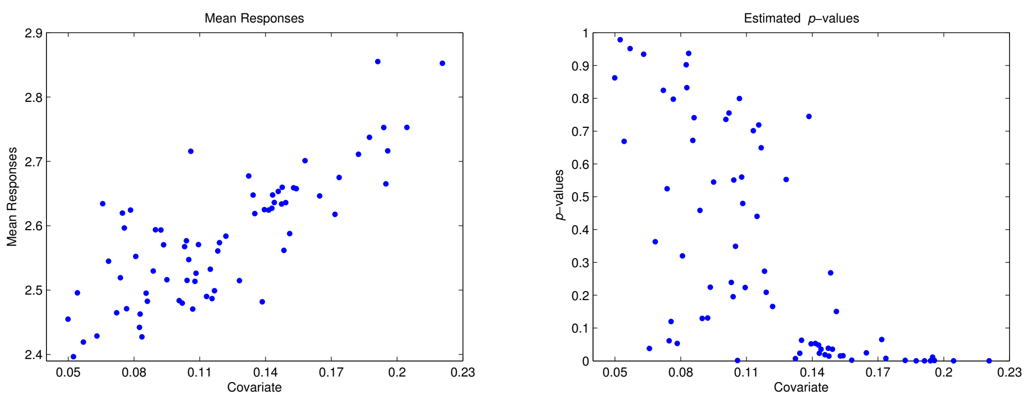



Figure 3.2: Plot of the average responses $\bar{Y}_i$ (left panel) and the estimated $p$-values (right panel) for the data from the queuing system.

previous sub-section yields the final estimates to be $\hat{d}_{m,n} = 0.1165$ and $\hat{\tau} = 2.5230$, assuming homoscedastic errors. The estimated $p$-values are plotted in the right panel of Figure 3.5.2 which illustrates the dichotomy in the behavior of the $p$-values – they are uniformly distributed to the left of $\hat{d}_{m,n}$, and close to zero beyond $\hat{d}_{m,n}$. Taking $k$ to be 1, and using the methodology described in the previous sub-section, the 90% and 95% CIs for the threshold turn out to be $[0.1051, 0.1276]$ and $[0.1031, 0.1301]$, respectively. Also, the adaptive upper 90% and 95% confidence bound for the threshold turn out to be 0.1348 and 0.1371, respectively. From the system's operator point of view the average delay of jobs exhibits a markedly increasing trend beyond a loading of 13%.

## 3.6 Conclusion

We conclude with a discussion of some open problems that can provide avenues for further investigation into this problem.

*Adaptivity.* We have provided a comprehensive treatment of the asymptotics of a $p$-value based procedure to estimate the threshold $d_0$ at which an unknown regression function $\mu$ takes off from its baseline value, with the aim of constructing CIs for $d_0$. We have assumed knowledge of the order of the 'cusp' of $\mu$ at $d_0$, which we need to achieve the optimal rate of convergence (and construct the corresponding CIs), though not for consistency. When $k$ is unknown, we have been able to construct adaptive one-sided CIs. However, constructing two-sided adaptive CIs remains a hard open problem and will be a topic of future research.

*Resampling.* A natural alternative to using the limit distribution (with estimated nuisance parameters) to construct CIs for $d_0$ would be to use bootstrap/resampling methods. Drawing from results obtained in similar change-point and non-standard problems (see e.g., Sen et al. (2010); Seijo and Sen (2011)) it is very likely that the usual bootstrap method will be inconsistent in our setup. However, model based bootstrap procedures have recently been studied in the change-point context and have been shown to work (Seijo and Sen, 2011). Similar ideas may work for our problem as well, but a thorough understanding of such bootstrap procedures is beyond the scope of this chapter. Subsampling has been proven to be consistent in our setting, but its finite sample properties were seen to be rather dismal.

# CHAPTER 4

# Asymptotics for the standard regression setting

In this chapter, we address the *inference problem* for the $p$-value procedure in the standard regression or simply the regression setting, i.e, we study the asymptotic properties of $\hat{d}_n$ arising out of a variant of criterion (2.4) from Chapter 2 and use these results to construct asymptotically valid CIs for the threshold, both in simulation settings and for two key motivating examples from Chapter 2. The problem, which falls within the sphere of non-regular M-estimation is rather hard, and involves non-trivial applications of techniques from modern empirical processes, as well as results from martingale theory and the theory of Gaussian processes. Along the way, we also deduce results on the large sample behavior of a kernel estimator at local points (see Lemma 4.4 and Proposition 4.13) that are of independent interest. In most of the literature, kernel estimates are considered at various fixed points and are asymptotically independent (Csörgő and Mielniczuk, 1995a,b; Robinson, 1997). Hence, they do not admit a functional limit. However, these estimates, when considered at local points, deliver an invariance principle; see Lemma 4.4 and the proof of Proposition 4.13.

As in the dose-response setting from Chapter 3, we show that the smoothness of the function in the vicinity of $d_0$ determines the rate of convergence of our estimator: for a "cusp" of order $k$ at $d_0$, the best possible rate of convergence turns out to be $n^{-1/(2k+1)}$. The limiting distribution of an appropriately normalized version of the estimator is that of the minimizer of the integral of a transformed Gaussian process. The limiting process is new, and while the uniqueness of the minimizer remains unclear (and appears to be a interesting nontrivial exercise in probability), we can bypass the lack of uniqueness and still provide a thorough mathematical framework to construct honest CIs. Under the assumption of uniqueness, which appears to be a very reasonable conjecture based on extensive simulations, we establish auxiliary results to construct asymptotically exact CIs as well.

The chapter is organized thus: we briefly recall the variant of the estimation procedure that we study in a fixed design setting and state the core assumptions in Section 4.1. In a fixed design, the resulting kernel estimates can be shown to be $m$-dependent, a feature that helps us in establishing the rate of convergence. The rate of convergence and the asymptotic distribution of the estimated threshold, along with some auxiliary results for constructing CIs, are deduced in Sections 4.2.1 and 4.2.2 assuming *a known* $\tau_0$. Asymptotic results for the other variants of the procedure are discussed in Section 4.2.3 and extensions of these results to the situation with *an unknown* $\tau_0$ are presented in Section 4.3. We study the coverage performance of the resulting CIs through simulations in Section 4.4. The applicability of our approach to short-range dependent data is the content of Section 4.5. We implement our procedure to two data examples in Section 4.6. Some concluding remarks are drawn in Section 4.7. The proofs of several technical results are available in Appendix C.

## 4.1 The Method

Consider the uniform fixed design regression model of the form:

$$Y_i = \mu\left(\frac{i}{n}\right) + \epsilon_i, \qquad 1 \leq i \leq n, \tag{4.1}$$

with $\epsilon_i$s i.i.d. having variance $\sigma_0^2$. Although we suppress the dependence on $n$, $Y_i$ and $\epsilon_i$ must be viewed as triangular arrays. Let $K$ be a symmetric probability density (kernel) and $h_n = h_0 n^{-\lambda}$ denote the smoothing bandwidth, for some $\lambda \in (0,1), h_0 > 0$. With a slightly different notation from Chapter 2, let

$$\hat{\mu}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} Y_i K\left(\frac{x - i/n}{h_n}\right) \tag{4.2}$$

denote an estimate of $\mu$. Let

$$\mathbb{M}_n(d) = \frac{1}{n} \sum_{i=1}^{n} \left[\Phi\left(\sqrt{nh_n}\left(\hat{\mu}\left(\frac{i}{n}\right) - \tau_0\right)\right) - \gamma\right] 1\left(\frac{i}{n} \leq d\right). \tag{4.3}$$

Then, an estimate of $d_0$, based on a non-normalized $p$-value criterion as in (2.4) of Chapter 2, is given by

$$\hat{d}_n = \operatorname*{sargmin}_{d \in [0,1]} \mathbb{M}_n(d),$$

where sargmin denotes the smallest argmin of the criterion function as earlier. Analyzing this method is useful in illustrating the core ideas while avoiding some of the tedious details encountered in analyzing the normalized $p$-value based estimate.

**Remark 4.1.** *As in Chapter 3, we first study the above method assuming a known $\tau_0$. When $\tau_0$ is unknown, a plug-in estimate can be substituted in its place (more about this in Section 4.3). Also, for any choice of $\gamma \in (1/2, 1)$ in (4.3), the estimator of $d_0$ is consistent. The proof follows along the lines of arguments for the proof of Theorem 2.2 in Chapter 2.*

Throughout this chapter, we make the following assumptions.

1. Assumptions on $\mu$:

   (a) $\mu$ is continuous on $[0, 1]$. We additionally assume that $\mu$ is Lipschitz continuous of order $\alpha_1$ with $\alpha_1 \in (1/2, 1]$.

(b) $\mu$ has a cusp of order $k$, $k$ being a *known* positive integer, at $d_0$, i.e., $\mu^{(l)}(d_0) = 0, 1 \leq l \leq k-1$, and $\mu^{(k)}(d_0+) > 0$, where $\mu^{(l)}(\cdot)$ denotes the $l$-th derivative of $\mu$. Also, the $k$-th derivative, $\mu^{(k)}(x)$ is assumed to be continuous and bounded for $x \in (d_0, d_0 + \zeta_0]$ for some $\zeta_0 > 0$.

2. The error $\epsilon$ possesses a continuous positive density on an interval.

3. Assumptions on the kernel $K$:

   (a) $K$ is a symmetric probability density.

   (b) $K(u)$ is non-increasing in $|u|$.

   (c) $K$ is compactly supported, i.e., $K(x) = 0$ when $|x| \geq L_0$, for some $L_0 > 0$.

   (d) $K$ is Lipschitz continuous of order $\alpha_2$ with $\alpha_2 \in (1/2, 1]$.

As a consequence of these assumptions, $\mu$ and $K$ are bounded, $\bar{K^2} = \int K^2(u)du < \infty$ and $E|W|^k < \infty$, where $W$ has density $K$. Also, both $\mu$ and $K$ are Lipschitz continuous of order $\alpha = \min(\alpha_1, \alpha_2)$. These facts are frequently used in the chapter. Common kernels such as the Epanechnikov kernel and the triangular kernel conveniently satisfy the assumptions mentioned above. The results in the next section are developed for a $\gamma \in (1/2, 1)$ (cf. Remark 4.1) and a known $\tau_0$. It will be seen in Section 4.3 that $\tau_0$ can be estimated at a sufficiently fast rate; consequently, even if $\tau_0$ is unknown, appropriate estimates can be substituted in its place to construct the $p$-values that are instrumental to the methods of this chapter, without changing the limit distributions. Without loss of generality, we take $\tau_0 \equiv 0$ in the next section, as one can work with $(Y_i - \tau_0)$s in place of $Y_i$s.

*Comparison with existing approaches.* Under the above assumptions, $d_0$ is a 'change-point' in the $k$-th derivative of $\mu$. Our procedure for estimating this change-point relies on the discrepancy of $p$-values, the construction of which requires a kernel-smoothed estimate (or if one desires, local polynomial estimate) of $\mu$. The estimation of a change-point in the derivative of a regression function has been studied by a number of authors using kernel-based strategies. However, the approaches in these papers are quite different from ours and more importantly, our problem cannot be solved by these methods without making stronger model assumptions than those above. In Müller (1992), the change-point is obtained by direct estimation of the $k$-th derivative ($k$ corresponds to $\nu$ in that paper) on either side of the change-point via one-sided kernels and measuring the difference between these estimates. In contrast, our approach does not rely on derivative estimation. We use an ordinary kernel function to construct a smooth estimate of $\mu$ which is required for the point wise testing procedures that lead to the $p$-values. In fact, a consistent estimate that attains the same rate of convergence as our current estimate could have been constructed using a simple regressogram estimator with an appropriate bin-width, in contrast to the approach in Müller (1992) which uses a $k$–times differentiable kernel. Müller (1992) also assumes that the $k$-th derivative of the regression function is at least twice continuously differentiable at all points except $d_0$ – see, pages 738–739 of that paper – which is stronger than our continuity assumption on $\mu^{(k)}$ (1(b) above). Cheng and Raimondo (2008) develop kernel methods for optimal estimation of the first derivative building on an idea by Goldenshluger et al. (2006), which is followed up in the context of dependent errors by Wishart and Kulik (2010), and Wishart (2009), but these papers do not consider the case $k > 1$. We also note that our method is fairly simple to implement.

## 4.2 Main Results

We consider the model stated in (4.1) with homoscedastic errors and uniform fixed design, and study the limiting behavior of $\hat{d}_n$ which minimizes (4.3). Results on the variant of the procedure discussed in Chapter 2 follow analogously and are stated in Section 4.2.3.

### 4.2.1 Rate of convergence

We start by fixing some notations. Let the variance of the statistic $\sqrt{nh_n}\hat{\mu}(x)$ be denoted by

$$\Sigma_n^2(x) = \Sigma_n^2(x, \sigma_0) = \mathrm{Var}(\sqrt{nh_n}\hat{\mu}(x)) = \frac{\sigma_0^2}{nh_n}\sum_{i=1}^{n} K^2\left(\frac{x - i/n}{h_n}\right). \tag{4.4}$$

Note that this converges to $\Sigma^2(x) = \sigma_0^2\bar{K^2}$ with $\bar{K^2} = \int K^2(u)du$. We first consider the population equivalent of $\mathbb{M}_n$, given here by $M_n(d) = E\left\{\mathbb{M}_n(d)\right\}$, and study the behavior of its smallest argmin. Recall that $\tau_0$ is taken to be zero without loss of generality. Let

$$Z_{in} = \frac{1}{\sqrt{nh_n}}\sum_{l=1}^{n}\epsilon_l K\left(\frac{i/n - l/n}{h_n}\right),$$

for $i = 1, \ldots, n$, and $Z_0$ be a standard normal random variable independent of $Z_{in}$'s. Also, let

$$\bar{\mu}(x) = \frac{1}{nh_n}\sum_{l=1}^{n}\mu\left(\frac{l}{n}\right)K\left(\frac{x - l/n}{h_n}\right). \tag{4.5}$$

Note that $\sqrt{nh_n}\hat{\mu}(i/n) = \sqrt{nh_n}\bar{\mu}(i/n) + Z_{in}$ and $\mathrm{Var}(Z_{in}) = \Sigma^2(i/n)$ with $\Sigma(\cdot)$ as in (4.4). We have

$$\begin{aligned} E\left[\Phi\left(\sqrt{nh_n}\hat{\mu}(i/n)\right)\right] &= E\left[\Phi\left(\sqrt{nh_n}\bar{\mu}(i/n) + Z_{in}\right)\right] \\ &= E\left[1\left(Z_0 \leq \sqrt{nh_n}\bar{\mu}(i/n) + Z_{in}\right)\right] \\ &= \Phi_{i,n}\left(\frac{\sqrt{nh_n}\bar{\mu}(i/n)}{\sqrt{1 + \Sigma_n^2(i/n)}}\right), \end{aligned} \tag{4.6}$$

where $\Phi_{i,n}$ denotes the distribution function of $(Z_0 - Z_{in})/\sqrt{1 + \Sigma_n^2(i/n)}$. Hence,

$$M_n(d) = \frac{1}{n}\sum_{i=1}^{n}\left\{\Phi_{i,n}\left(\frac{\sqrt{nh_n}\bar{\mu}(i/n)}{\sqrt{1 + \Sigma_n^2(i/n)}}\right) - \gamma\right\}1\left(\frac{i}{n} \leq d\right).$$

For $L_0 h_n \le i/n \le 1 - L_0 h_n$, $\Phi_{i,n}$'s and $\Sigma_n(i/n)$'s do not vary with $i$. We denote them by $\tilde{\Phi}_n$ and $\tilde{\Sigma}_n$ for convenience. Using Corollary C.2 and (C.1) from Appendix C, $\tilde{\Sigma}_n$ converges to $\sigma_0\sqrt{\bar{K^2}}$. Also, for such $i$'s, any $\eta > 0$ and sufficiently large $n$,

$$\frac{1}{nh_n\tilde{\Sigma}_n^2}\sum_{l:|l-i|\le L_0 nh_n}^{n} E\left[\epsilon_l^2 K^2\left(\frac{(i-l)/n}{h_n}\right) 1\left(\frac{|\epsilon_l| K\left((i-l)/(nh_n)\right)}{\sqrt{nh_n}\tilde{\Sigma}_n(l/n)} > \eta\right)\right]$$

is bounded by

$$\frac{2\left\lceil 2L_0 nh_n\right\rceil \|K\|_\infty^2}{nh_n(\sigma_0^2\bar{K^2})} E\left[\epsilon_1^2 1\left(\frac{2\|K\|_\infty}{nh_n(\sigma_0\sqrt{\bar{K^2}})}|\epsilon_1| > \eta\right)\right],$$

which converges to zero. Hence, by Lindeberg–Feller CLT, $Z_{in}/\tilde{\Sigma}_n$ and consequently, $\tilde{\Phi}_n$ converge weakly to $\Phi$. In fact, for any $i$, we can also show that $\Phi_{i,n}$ converges weakly to $\Phi$.

Let $d_n = \operatorname{sargmin}_d M_n(d)$. As mentioned earlier, sargmin denotes the smallest argmin of the objective function $M_n$ which does not have a unique minimizer. The following lemma provides the rate at which $d_n$ converges to $d_0$.

**Lemma 4.2.** *Let* $\nu_n = \min(h_n^{-1}, (nh_n)^{1/2k})$. *Then* $\nu_n(d_n - d_0) = O(1)$.

*Proof.* It can be shown by arguments analogous to proof of Theorem 2.2 from Chapter 2 that $(d_n - d_0)$ is $o(1)$. As $d_0$ is an interior point of [0,1], $d_n \in (L_0 h_n, 1 - L_0 h_n)$ and corresponds to a local minima of $M_n$ for sufficiently large $n$, i.e., $d_n$ satisfies

$$\tilde{\Phi}_n\left(\frac{\sqrt{nh_n}\bar{\mu}(d_n)}{\sqrt{1+\tilde{\Sigma}_n^2}}\right) \le \gamma \text{ and } \tilde{\Phi}_n\left(\frac{\sqrt{nh_n}\bar{\mu}(d_n+1/n)}{\sqrt{1+\tilde{\Sigma}_n^2}}\right) > \gamma. \tag{4.7}$$

By Pólya's theorem, $\tilde{\Phi}_n$ converges uniformly to $\Phi$. Consequently,

$$0 \le \sqrt{nh_n}\bar{\mu}(d_n) \le \Phi^{-1}(\gamma)\sqrt{1+\sigma_0^2\bar{K^2}} + o(1). \tag{4.8}$$

Note that $\bar{\mu}(x) = 0$ for $x < d_0 - L_0 h_n$ and $\tilde{\Phi}_n(0)$ converges to $\Phi(0) = 0.5 < \gamma$. So, if $d_n < d_0$, then for (4.7) to hold, $d_n + 1/n + L_0 h_n > d_0$ for large $n$ and thus $h_n^{-1}(d_n - d_0) = O(1)$ which gives the result. Also, when $d_0 < d_n \leq d_0 + L_0 h_n$, the result automatically holds. So, it suffices to consider the case $d_n > d_0 + L_0 h_n$.

Let $u_n(x, v) = (1/h_n)\mu(v) K((x - v)/h_n)$ for $x \in [0, 1]$ and $v \in \mathbb{R}$. By Lemma C.1 from Appendix C,

$$\left| \bar{\mu}(d_n) - \int_0^1 u_n(d_n, v) dv \right| = O\left( \frac{1}{(nh_n)^\alpha} \right).$$

By a change of variable, $\int_0^1 u_n(d_n, v)dv = \int_{-L_0}^{L_0} \mu(d_n + uh_n)K(u)du$ for large $n$. As $d_n > d_0 + L_0 h_n$, the first part of the integrand, $\mu(d_n + uh_n)$, is positive for $u \in [-L_0, L_0]$. Let $[-L_1, L_1]$ be an interval where $K$ is positive. Such an interval exists due to assumptions 4(a) and 4(b). Hence, $\int_{-L_1}^{L_1} \mu(d_n + uh_n)K(u)du = 2L_1 \mu(d_n + \xi_n h_n)K(\xi_n) \leq \int u_n(d_n, v)dv$, where $\xi_n$ is some point in $[-L_1, L_1]$. Using Taylor expansion around $d_0$, $\mu(d_n + \xi_n h_n) = \left\{ \mu^{(k)}(\zeta_n)/k! \right\} (d_n + \xi_n h_n - d_0)^k$, for some $\zeta_n$ lying between $d_0$ and $d_n + \xi_n h_n$. By (4.8), we get

$$2L_1 \frac{\mu^{(k)}}{k!}(\zeta_n)(d_n + \xi_n h_n - d_0)^k K(\xi_n) \quad = \quad O((nh_n)^{-1/2}).$$

As $d_n \to d_0$, $\mu^{(k)}(\zeta_n)$ converges to $\mu^{(k)}(d_0+)$, which is positive. Also, as $\xi_n \in [-L_1, L_1]$, $K(\xi_n)$ is bounded away from zero, and thus $(d_n + \xi_n h_n - d_0) = O((nh_n)^{-1/2k})$, which yields the result. $\square$

As $\hat{d}_n$ is, in fact, estimating $d_n$, its rate of convergence for $d_0$ can at most be $\nu_n^{-1}$. Fortunately, $\nu_n^{-1}$ turns out to be the exact rate of convergence of $\hat{d}_n$.

**Theorem 4.3.** *Let $\nu_n$ be as defined in Lemma 4.2. Then $\nu_n(\hat{d}_n - d_0) = O_p(1)$.*

The proof is given in Section C.1 of Appendix C. It involves coming up with an appropriate distance $\rho_n$ based on the behavior of $M_n$ near $d_0$ (Lemma C.3) and then establishing a modulus of continuity bound for $\mathbb{M}_n - M_n$ with respect to $\rho_n$. As the summands that constitute $\mathbb{M}_n$ are dependent, the latter cannot be handled directly through VC or bracketing results (Theorems 2.14.1 or 2.14.2 of van der Vaart and Wellner (1996)); rather, we require a blocking argument followed by an application of Doob's inequality to the blocks.

The optimal rate is attained when $h_n^{-1} \sim (nh_n)^{1/(2k)}$ and corresponds to $h_n = h_0 n^{-1/(2k+1)}$ and $\nu_n = n^{1/(2k+1)}$. We now deduce the asymptotic distribution for this particular choice of bandwidth.

### 4.2.2 Asymptotic Distribution

With $h_n = h_0 n^{-1/(2k+1)}$, we study the limiting behavior of the process

$$Z_n(t) = h_n^{-1} \left[ \mathbb{M}_n(d_0 + th_n) - \mathbb{M}_n(d_0) \right], \quad t \in \mathbb{R}, \tag{4.9}$$

where $\mathbb{M}_n$ is defined in (4.3). The process $Z_n(t)$ is minimized at $h_n^{-1}(\hat{d}_n - d_0)$. At the core of the process $Z_n(t)$ lies the estimator $\hat{\mu}$, computed at local points $d_0 + th_n$. Let

$$W_n(t) = \sqrt{nh_n}\hat{\mu}(d_0 + th_n) \tag{4.10}$$

and $B_{loc}(\mathbb{R})$ denote the space of locally bounded functions on $\mathbb{R}$, equipped with the topology of uniform convergence on compacta. We have the following lemma on the limiting behavior of $W_n$.

**Lemma 4.4.** *There exists a Gaussian process $W(t), t \in \mathbb{R}$, with almost sure continuous paths and drift*

$$m(t) = E(W(t)) = \frac{h_0^{k+1/2} \mu^{(k)}(d_0+)}{k!} \int_{-\infty}^{t} (t-v)^k \, K(v) \, dv$$

*and covariance function $Cov(W(t_1), W(t_2)) = \sigma_0^2 \int K(t_1 + u) K(t_2 + u) du$ such that* *the process $W_n(\cdot)$ converges weakly to $W(\cdot)$ in $B_{loc}(\mathbb{R})$.*

The proof is given in Section C.2 of Appendix C. For brevity, $-\int_y^x$ is written as $\int_x^y$ whenever $x > y$.

**Theorem 4.5.** *For $h_n = h_0 n^{-1/(2k+1)}$ and $t \in \mathbb{R}$, the process $Z_n(t)$ converges weakly to the process $Z(t) = \int_0^t [\Phi(W(y)) - \gamma]\, dy$ in $B_{loc}(\mathbb{R})$.*

*Proof.* Split $Z_n(t)$ as $I_n(t) + II_n(t)$, where

$$\begin{aligned} I_n(t) &= \frac{1}{nh_n}\sum_{i=1}^{n}\left[\left\{\Phi\left(\sqrt{nh_n}\hat{\mu}(i/n)\right) - \gamma\right\}\right. \\ &\quad \times \left.\left\{1\left(\frac{i}{n} \leq d_0 + th_n\right) - 1\left(\frac{i}{n} \leq d_0\right)\right\}\right] \\ &\quad -\frac{1}{h_n}\int_{d_0}^{d_0+th_n}\left(\Phi\left(\sqrt{nh_n}\hat{\mu}(x)\right) - \gamma\right) dx \end{aligned}$$

and $II_n = h_n^{-1}\int_{d_0}^{d_0+th_n}\left(\Phi\left(\sqrt{nh_n}\hat{\mu}(x)\right) - \gamma\right) dx$. Fix $T > 0$ and let $t \in [-T, T]$. Using arguments almost identical to those for proving Lemma C.1 in Appendix C, we have

$$\begin{aligned} |I_n(t)| &\leq \sum_{\substack{|d_0 - i/n| \\ \leq Th_n}} \int_{i/n}^{(i+1)/n} \frac{1}{h_n}\left|\Phi\left(\sqrt{nh_n}\hat{\mu}(i/n)\right) - \Phi\left(\sqrt{nh_n}\hat{\mu}(x)\right)\right| dx \\ &\quad + O\left(\frac{1}{nh_n}\right) + \frac{\gamma}{nh_n}(\lfloor n(d_0 + th_n)\rfloor - \lfloor n(d_0)\rfloor) - \gamma t, \end{aligned}$$

where the $O(1/(nh_n))$ factor accounts for the boundary terms. Using the fact that $x - 1 \leq \lfloor x \rfloor \leq x + 1$, the term $(\gamma/(nh_n))(\lfloor n(d_0 + th_n)\rfloor - \lfloor n(d_0)\rfloor) - \gamma t$ is bounded by $2\gamma(1/(nh_n) + T/n)$ which goes to zero. The sum of integrals in the above display is further bounded by

$$
\begin{aligned}
&\frac{\lceil 2Tnh_n \rceil}{nh_n} \sup_{\substack{|x-y|<1/n \\ x,y\in[d_0-Th_n, d_0+Th_n]}} \left| \Phi\left(\sqrt{nh_n}\hat{\mu}(x)\right) - \Phi\left(\sqrt{nh_n}\hat{\mu}(y)\right) \right| \\
&\qquad \leq \frac{\lceil 2Tnh_n \rceil}{2\pi nh_n} \sup_{\substack{|u-v|<1/(nh_n) \\ u,v\in[-T,T]}} |W_n(u) - W_n(v)| \,.
\end{aligned}
$$

The above display goes in probability to zero due to the asymptotic equicontinuity of the process $W_n$ and hence the term $I_n$ converges in probability to zero uniformly in $t$ over compact sets. Further, we have

$$
\begin{aligned}
II_n(t) &= h_n^{-1} \int_{d_0}^{d_0+th_n} \left(\Phi\left(\sqrt{nh_n}\hat{\mu}(x)\right) - \gamma\right) dx \\
&= \int_0^t \left[\Phi\left(\sqrt{nh_n}\hat{\mu}(d_0+yh_n)\right) - \gamma\right] dy \\
&= \int_0^t \left[\Phi\left(W_n(y)\right) - \gamma\right] dy.
\end{aligned}
$$

As the mapping $W(\cdot) \mapsto \int_0^{\cdot} \Phi(W(y))dy$ from $B_{loc}(\mathbb{R})$ to $B_{loc}(\mathbb{R})$ is continuous, using Lemma 4.4, the term $II_n$ converges weakly to the process $\int_0^t [\Phi(W(y)) - \gamma]\, dy$, $t \in \mathbb{R}$. This completes the proof. $\square$

A conservative asymptotic CI for $d_0$ can be obtained using the following result.

**Theorem 4.6.** *The process $Z(t)$ goes to infinity almost surely (a.s.) as $|t| \to \infty$. Moreover, let $\xi_0^s$ and $\xi_0^l$ denote the smallest and the largest minimizers of the process $Z$. Also, let $c_{\alpha/2}^s$ and $c_{1-\alpha/2}^l$ be the $(\alpha/2)$th and $(1-\alpha/2)$th quantiles of $\xi_0^s$ and $\xi_0^l$ respectively. For $h_n = h_0 n^{-1/(2k+1)}$, we have*

$$
\liminf_{n\to\infty} P[c_{\alpha/2}^s < h_n^{-1}(\hat{d}_n - d_0) < c_{1-\alpha/2}^l] \geq 1 - \alpha.
$$

Note that $\xi_0^s$ and $\xi_0^l$ are indeed well defined by continuity of the sample paths of $Z$ and the fact that $Z(t)$ goes to infinity as $|t| \to \infty$. Also, they are Borel measurable as, say for $\xi_0^s$, the events $[\xi_0^s \leq a]$ and the measurable event $[\inf_{t \leq a} Z(t) \leq \inf_{t > a} Z(t)]$ are equivalent for any $a \in \mathbb{R}$. Hence $c^s_{\alpha/2}$ and $c^l_{1-\alpha/2}$ are well defined. The proof of the result is given in Section C.3 of Appendix C.

A minimum of the underlying limiting process lies in the set $\{y : \Phi(W(y)) = \gamma\}$. As any fixed number has probability zero of being in this set, the distributions of $\xi_0^s$ and $\xi_0^l$ are continuous. The process $\{W(y) : y \in \mathbb{R}\}$ has zero drift for $y < -L_0$ and is therefore stationary to the left of $-L_0$. Hence, it must cross $\gamma$ infinitely often implying that $Z$ has multiple local extrema. On the other hand, simulations *strongly* suggest that $Z$ has a unique argmin though a theoretical justification appears intractable at this point. The issue of the uniqueness of the argmin of a stochastic process has mostly been addressed in context of Gaussian processes (Lifshits, 1982; Kim and Pollard, 1990; Ferger, 1999), certain transforms of compound Poisson processes (Ermakov, 1976; Pflug, 1983) and set-indexed Brownian motion (Müller and Song, 1996). These techniques do not apply to our setting; in fact, an analytical justification of the uniqueness of the minimizer of $Z$ appears non-trivial. As the simulations provide strong evidence in support of a unique argmin, we use the following result for constructing CIs in practice.

**Theorem 4.7.** *Assuming that the process $Z$ has a unique argmin, we have*

$$h_n^{-1}(\hat{d}_n - d_0) \xrightarrow{d} \operatorname*{argmin}_{t \in \mathbb{R}} \{Z(t)\},$$

*for $h_n = h_0 n^{-1/(2k+1)}$.*

**Remark 4.8.** *As deduced in Chapter 3, the limit distribution in the dose-response setting is governed by the minimizer of a generalized compound Poisson process, in contrast to the integral of a transformed Gaussian process that appears in the stan-*

*dard regression setting. The appearance of a markedly different transformed Gaussian process is an outcome of the local spatial averaging of responses needed to construct the p-values in the absence of multiple replications.*

Note that when the argmin is unique, Theorem 4.6 and Theorem 4.7 yield the same CI. The proof of Theorem 4.7 is a direct application of the argmin(argmax)-continuous mapping theorem; see Kim and Pollard (1990, Theorem 2.7) or van der Vaart and Wellner (1996, Theorem 3.2.2).

### 4.2.3 Limit distributions for variants of the procedure

The rates of convergence and asymptotic distributions can be obtained similarly for most of the variants of the procedure that were discussed in Chapter 2. In what follows, we state the limiting distributions for some of these variants.

Results analogous to Theorem 4.6 can be shown to hold in the setting with heteroscedastic errors, i.e., $\mathrm{Var}(\epsilon_i) = \sigma_0^2(i/n)$, where $\sigma_0^2(\cdot)$ is a positive continuous function. The process $Z$ has the same form as in Theorem 4.6 apart from the fact that the $\sigma_0^2$ involved in the covariance kernel of the process $W$ that appears in the definition of $Z$ is replaced by $\sigma_0^2(d_0)$. When normalized $p$-values are used to estimate $d_0$, we have the following form for the limiting distribution; an outline of its proof is given in Section C.4 of Appendix C.

**Proposition 4.9.** *Consider the setting with homoscedastic errors and covariates sampled from the fixed uniform design, as discussed in Section 4.1. Let $\hat{\sigma}$ be an estimate of $\sigma_0$ such that $\sqrt{n}(\hat{\sigma} - \sigma_0) = O_p(1)$ and let $\hat{d}_n^1$ denote the estimate obtained from minimizing*

$$\begin{aligned} \widetilde{\mathbb{M}}_n(d) &\equiv \widetilde{\mathbb{M}}_n(d, \hat{\sigma}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\{ \Phi\left( \frac{\sqrt{nh_n}(\hat{\mu}(i/n) - \tau_0)}{\Sigma_n(i/n, \hat{\sigma})} \right) - \gamma \right\} 1\left( \frac{i}{n} \leq d \right). \end{aligned}$$

*Let $h_n = h_0 n^{-1/(2k+1)}$ and $W^1(t), t \in \mathbb{R}$, be a Gaussian process with drift*

$$E(W^1(t)) = \frac{h_0^{k+1/2}\mu^{(k)}(d_0+)}{k!\sigma_0\sqrt{\bar{K^2}}} \int\limits_{-\infty}^{t} (t-v)^k \, K\,(v)\, dv$$

*and covariance function $Cov(W^1(t_1), W^1(t_2)) = (\bar{K^2})^{-1} \int K(t_1+u)K(t_2+u)du$. Let $Z^1(t) = \int_0^t \{\Phi(W^1(y)) - \gamma\}dy$, for $t \in \mathbb{R}$. If $\hat{\sigma}$ is a $\sqrt{n}$-consistent estimate of $\sigma_0$, then $h_n^{-1}(\hat{d}_n^1 - d_0)$ is $O_p(1)$. For $Z^1$ possessing a unique argmin a.s., we have*

$$h_n^{-1}(\hat{d}_n^1 - d_0) \xrightarrow{d} \operatorname*{argmin}_{t\in\mathbb{R}} Z^1(t).$$

When the covariate is sampled from a random design with heteroscedastic errors, the result extends as follows for the estimate based on non-normalized $p$-values. Recall from Chapter 2 that such an estimator of $d_0$ has the form

$$\tilde{d}_n = \operatorname*{sargmin}_{d\in(0,1)} \mathbb{P}_n \left[\left\{\Phi\left(\sqrt{nh_n}(\tilde{\mu}(X) - \tau_0)\right) - \gamma\right\} 1(X \le d)\right], \tag{4.11}$$

where $\tilde{\mu}$ is the Nadaraya-Watson estimator, i.e.,

$$\tilde{\mu}(x) = \frac{\mathbb{P}_n[YK\,((x-X)/h_n)]}{\mathbb{P}_n[K\,((x-X)/h_n)]}.$$

We have the following result on the limiting distribution of $\tilde{d}_n$.

**Proposition 4.10.** *Consider the setting with covariates sampled from a random design with design density $f$ and heteroscedastic errors. The variance function $\sigma_0^2(x) = Var(\epsilon \mid X = x)$ is assumed to be continuous and positive. Let $h_n = h_0 n^{-1/(2k+1)}$ and $\tilde{W}(t), t \in \mathbb{R}$, be a Gaussian process with drift*

$$E(\tilde{W}(t)) = \frac{h_0^{k+1/2}\mu^{(k)}(d_0+)}{k!} \int\limits_{-\infty}^{t} (t-v)^k \, K\,(v)\, dv$$

*and covariance function* $Cov(\tilde{W}(t_1), \tilde{W}(t_2)) = \frac{\sigma_0^2(d_0)}{f(d_0)} \int K(t_1 + u)K(t_2 + u)du$. *Let* $\tilde{Z}(t) = \int_0^t \{\Phi(\tilde{W}(y)) - \gamma\}dy$, *for* $t \in \mathbb{R}$. *For* $\tilde{d}_n$ *defined as in* (4.11), *assume that* $h_n^{-1}(\tilde{d}_n - d_0)$ *is* $O_p(1)$. *For* $\tilde{Z}$ *possessing a unique argmin a.s., we have*

$$h_n^{-1}(\tilde{d}_n - d_0) \xrightarrow{d} \underset{t \in \mathbb{R}}{\operatorname{argmin}}\, \tilde{Z}(t).$$

A sketch of the proof is given in Section C.5 of Appendix C.

## 4.3 The case of an unknown $\tau_0$

Although most of the results have been deduced under the assumption of a known $\tau_0$, in real applications $\tau_0$ is generally not known. In this situation, one would need to impute an estimate of $\tau_0$ in the objective function to carry out the procedure. It can be shown that the rate of convergence and the limit distribution does not change as long as we have a $\sqrt{n}$-consistent estimator of $\tau_0$. The following result makes this formal; its proof is given in Section C.6 of Appendix C.

**Proposition 4.11.** *Let* $\hat{d}_n$ *now denote the minimizer of*

$$\mathbb{M}_n(d, \hat{\tau}) = \frac{1}{n}\sum_{i=1}^{n}\left[\Phi\left(\sqrt{nh_n}\left(\hat{\mu}\left(\frac{i}{n}\right) - \hat{\tau}\right)\right) - \gamma\right] 1\left(\frac{i}{n} \le d\right),$$

*where* $\sqrt{n}(\hat{\tau} - \tau_0) = O_p(1)$ *and* $h_n = h_0 n^{-1/(2k+1)}$. *Then* $h_n^{-1}(\hat{d}_n - d_0)$ *is* $O_p(1)$. *Assuming that the process* $Z$ *defined in Theorem 4.5 has a unique argmin, we have*

$$h_n^{-1}(\hat{d}_n - d_0) \xrightarrow{d} \underset{t \in \mathbb{R}}{\operatorname{argmin}}\{Z(t)\}.$$

Quite a few choices are possible for estimating $\tau_0$. If $d_0$ can be safely assumed to be larger than some $\eta$, then a simple averaging of the observations below $\eta$ would yield a $\sqrt{n}$-consistent estimator of $\tau_0$. If a proper choice of $\eta$ is not available, one can obtain an initial (consistent) estimate of $\tau_0$ using the method proposed in Section 2.1.4 of Chapter 2 (see (4.12)), compute $\hat{d}_n$ and then average the responses from, say, $[0, c\hat{d}_n]$, $c \in (0,1)$, to obtain a $\sqrt{n}$-consistent estimator of $\tau_0$. This leads to an iterative procedure which we discuss in more detail in Section 4.4. In what follows, we justify that such an estimate of $\tau_0$ is indeed $\sqrt{n}$-consistent.

**Lemma 4.12.** *Let $0 < c < 1$. For any consistent estimator $d'_n$ of $d_0$, define*

$$\hat{\tau} := \frac{1}{\lfloor ncd'_n \rfloor} \sum_{i=1}^{n} Y_i 1\left(\frac{i}{n} \leq cd'_n\right).$$

*We have $\sqrt{n}(\hat{\tau} - \tau_0) = O_p(1)$.*

*Proof.* Note that for $T > 0$ and $0 < \kappa < \min(c, (1-c))d_0$,

$$\begin{aligned} P\left[\sqrt{n}|\hat{\tau} - \tau| > T\right] \;\; &\leq \;\; P\left[\sqrt{n}|\hat{\tau} - \tau| > T, \kappa < cd'_n < d_0 - \kappa\right] \\ &\quad + P\left[d'_n - d_0 < (\kappa - cd_0)/c\right] \\ &\quad + P\left[d'_n - d_0 > ((1-c)d_0 - \kappa)/c\right]. \end{aligned}$$

The second and the third term on the right side of the above display both converge to zero. Also,

$$\begin{aligned} &E\left[n(\hat{\tau} - \tau)^2 1\left(\kappa < cd'_n < d_0 - \kappa\right)\right] \\ &\quad = \;\; nE\left[\left(\frac{1}{\lfloor ncd'_n \rfloor} \sum_{i=1}^{\lfloor ncd'_n \rfloor} \epsilon_i\right)^2 1\left(\kappa < cd'_n < d_0 - \kappa\right)\right] \\ &\quad \leq \;\; \frac{n}{(n\kappa - 1)^2} E\left[\sup_{a \leq d_0 - \kappa} \left(\sum_{i=1}^{\lfloor na \rfloor} \epsilon_i\right)^2\right] \\ &\quad \leq \;\; \frac{4n}{(n\kappa - 1)^2} E\left[\left(\sum_{i=1}^{\lfloor n(d_0 - \kappa) \rfloor} \epsilon_i\right)^2\right] \leq \frac{4n(n(d_0 - \kappa) + 1)\sigma_0^2}{(n\kappa - 1)^2}. \end{aligned}$$

Here, the penultimate step followed from Doob's inequality. Hence,

$E\left[n(\hat{\tau}-\tau)^2 1\left(\kappa < cd_n' < d_0 - \kappa\right)\right] = O(1)$. Thus, by Chebyshev's inequality,

$$P\left[\sqrt{n}|\hat{\tau}-\tau| > T, \kappa < cd_n' < d_0 - \kappa\right] \leq O(1)/T^2$$

which can be made arbitrarily small by choosing $T$ large. This completes the proof.

□

## 4.4 Simulations

We consider three choices for the underlying regression function $\mu_k(x) = [2(x - 0.5)]^k 1(x > 0.5), x \in [0,1]$, $k = 1, 2$ and $\mu_3(x) = [(x - 0.5) + (1/5)\sin(5(x - 0.5)) + 0.3\sin(100(x - 0.5)^2)]1(x > 0.5)$. All these functions are at their baseline value 0 up to $d_0 = 0.5$. The functions $\mu_1$ (linear) and $\mu_2$ (quadratic) both rise to 1 while $\mu_3$ exhibits non-isotonic sinusoidal behavior after rising at $d_0$. The right derivative at $d_0$, a factor that appears in the limiting process $Z$, is the same for $\mu_1$ and $\mu_3$. The functions are plotted in the upper left panel of Figure 4.1. The functions $\mu_1$ and $\mu_2$ are paired up with normally distributed errors having mean 0 and standard deviation $\sigma_0 = 0.1$, while the noise added with $\mu_3$ is from a $t$-distribution with 5 degrees of freedom, scaled to have the standard deviation $\sigma_0$. The three models, $\mu_1$ with normal errors, $\mu_2$ with normal errors and $\mu_3$ with $t$-distributed errors, are referred to by the name of their regression functions only. We work with $\gamma = 3/4$ as extreme values of $\gamma$ (close to 0.5 or 1) tend to cause instabilities.

We construct the estimate of $d_0$ using the normalized $p$-values as they exhibit better finite sample performance and study the coverage performance of the approximate CIs obtained from the limiting distributions with estimated nuisance parameters. The error variance $\sigma_0^2$ is estimated in a straightforward manner using $\hat{\sigma}^2 = (1/n)\sum_i \{Y_i - \hat{\mu}(i/n)\}^2$. More sophisticated estimates of the error variance are also available (Gasser et al., 1986; Hall et al., 1990) but we avoid them for the sake of simplicity. We use the Epanechnikov kernel for constructing the estimate of $\mu$. For moderate samples, the bad behavior of kernel estimates near the boundary affects the coverage performance. In order to correct for this, we only consider the terms between $h_n$ to $1-h_n$ in our objective function, i.e., for $d \in (h_n, 1-h_n)$,

$$\mathbb{M}_n(d, \tau_0) = \frac{1}{n} \sum_{h_n \le \frac{i}{n} \le 1-h_n} \left\{ \Phi\left( \frac{\sqrt{nh_n}(\hat{\mu}(i/n) - \tau_0)}{\Sigma_n(i/n, \hat{\sigma})} \right) - \gamma \right\} 1\left( \frac{i}{n} \le d \right).$$

The asymptotic distribution of the minimizer of this restricted criterion function still has the same form as in *Proposition 4.9*. A good choice for $h_0$ in the optimal bandwidth $h_n = h_0 n^{-1/(2k+1)}$ can be obtained through minimizing the MSE of $\hat{\mu}(d_0)$. Standard calculations shown that

$$\begin{aligned}
\text{Bias}(\hat{\mu}(d_0)) &= \frac{\mu^{(k)}(d_0+)}{k!} h_n^k E[W^k 1(W > 0)] + o(h_n^k) + O\left( \frac{1}{(nh_n)^\alpha} \right), \text{ and} \\
\text{Var}(\hat{\mu}(d_0)) &= \frac{\sigma_0^2}{nh_n} \bar{K^2} + o\left( \frac{1}{nh_n} \right),
\end{aligned}$$

where $W$ has density $K$. The MSE is minimized at $h_n = h_0^{opt} n^{-1/(2k+1)}$ where

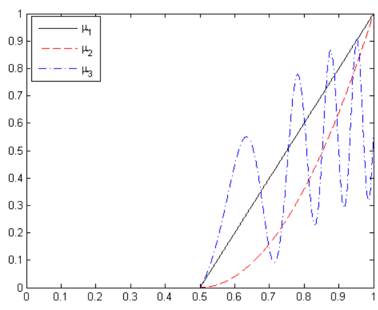
$$h_0^{opt} = \left[ \frac{\sigma_0^2 \bar{K^2} (k!)^2}{2k\{\mu^{(k)}(d_0+) E[W^k 1(W > 0)]\}^2} \right]^{-1/(2k+1)}.$$

This bandwidth goes to 0 at the right rate needed for estimating $d_0$. Moreover, efficient estimation of $\mu$ in the vicinity of $d_0$ is likely to aid in estimating $d_0$. Hence, we advocate the use of this choice of $h_0$ for our procedure.

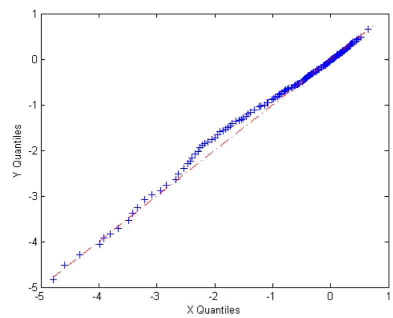
With the above mentioned choice of $h_0$, we compare the distribution of $h_n^{-1}(\hat{d}_n - d_0)$ for $n = 1000$ data points over 5000 replications with the deduced asymptotic distribution. As $\tau_0$ is assumed unknown, we implement an iterative scheme. We obtain an initial estimate of $\tau_0$ using the method prescribed in Chapter 2, i.e.,

$$\hat{\tau}_{init} = \underset{\tilde{\tau} \in \mathbb{R}}{\operatorname{argmin}} \sum \left\{ \Phi \left( \frac{\sqrt{n h_n} (\hat{\mu}(i/n) - \tilde{\tau})}{\Sigma_n(i/n, \hat{\sigma})} \right) - \frac{1}{2} \right\}^2 . \tag{4.12}$$

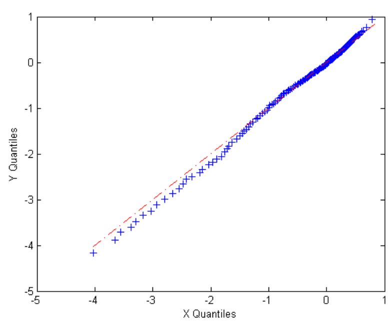
This estimate of $\tau_0$, based on $h_0^{opt}$, is used to compute $\hat{d}_n$. We re-estimate $\tau_0$ by averaging the responses for which $i/n \in [0, 0.9\hat{d}_n]$ and proceed thus. The Q-Q plots are shown in Figure 4.1 which show considerable agreement between the two distributions.



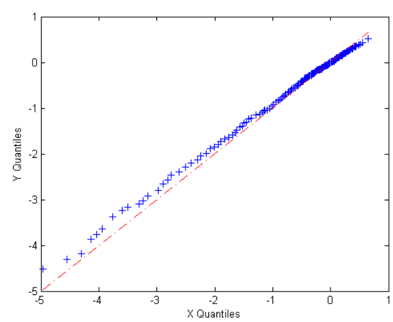

(a) Regression Functions




(b) Q-Q plot under $\mu_1$




(c) Q-Q plot under $\mu_2$




(d) Q-Q plot under $\mu_3$

Figure 4.1: The three regression functions and Q-Q plots for the normalized estimate $h_n^{-1}(\hat{d}_n - d_0)$ computed with $n = 1000$ over 5000 replications.

Next, we explore the coverage performance of the CIs constructed by imputing estimates of the nuisance parameters in the limiting distribution. Computing $h_0$ requires the knowledge of the $k$-th derivative of $\mu$ at $d_0$ which we also need to generate from the limit process. To estimate $\mu^{(k)}(d_0+)$, first observe that $\mu(x) = \mu^{(k)}(d_0+)(x - d_0)^k/k! + o((x - d_0)^k)$ for $x > d_0$. Hence, an estimate of $\mu^{(k)}(d_0+)$ can be obtained by fitting a $k$-th power of the covariate to the right of $\hat{d}_n$. More precisely, an estimate of $\xi_0 \equiv \mu^{(k)}(d_0+)/k!$ is given by

$$\begin{aligned} \hat{\xi} &= \underset{\xi}{\operatorname{argmin}} \sum_{i=1}^{n} \{Y_i - \xi(i/n - \hat{d}_n)^k\}^2 1(i/n \in (\hat{d}_n, \hat{d}_n + b_n]) \\ &= \frac{\sum Y_i (i/n - \hat{d}_n)^k 1(i/n \in (\hat{d}_n, \hat{d}_n + b_n])}{\sum (i/n - \hat{d}_n)^{2k} 1(i/n \in (\hat{d}_n, \hat{d}_n + b_n])}, \end{aligned}$$

where $b_n \downarrow 0$ and $nb_n^{2k+1} \to \infty$. For the optimal $h_n$, this provides a good estimate of $\xi_0$.

We include this in our iterative method where we start with an arbitrary choice of $h_0$ and compute $\hat{\tau}_{init}$. We use $\hat{\tau}_{init}$ to compute $\hat{d}_n$ and $\hat{\mu}^{(k)}(d_0+)$. The parameter $\hat{\mu}^{(k)}(d_0+)$ is estimated using a reasonably wide smoothing bandwidth $b_n$, $b_n = 5(n/\log n)^{-1/(2k+1)}$. These initial estimates are used to compute the next level estimate of $h_0$ using the expression for $h_0^{opt}$. We re-estimate $\tau_0$ by averaging the responses for which $i/n \in [0, 0.9\hat{d}_n]$ and proceed thus. On average, the estimates stabilize within 7 iterations. The coverage performance over 5000 replications is given below in Table 4.1. The approximate CIs mostly exhibit over-coverage for moderate sample sizes for $\mu_1$ and $\mu_3$ but converge to the desired confidence levels for large $n$. Also, the limiting distribution is same under models $\mu_1$ and $\mu_3$ which is evident from the coverages and the length of CIs for large $n$.

## 4.5 Dependent data

We briefly discuss the extension to dependent data in this section. Our problem is relevant to applications from time series models (see Section 4.6) where it is not reasonable to assume that the errors $\epsilon_i$'s are independent. A model of the form (4.1) can be assumed here with the exception that the errors now arise from a stationary sequence $\{\ldots \epsilon_{-1}, \epsilon_0, \epsilon_1, \ldots\}$ and exhibit short-range dependence in the sense of

| $n$ | 90% CI | | 95% CI | |
|---|---|---|---|---|
| | T | E | T | E |
| 30 | 0.949 (0.462) | 0.961 (0.614) | 0.989 (0.588) | 0.987 (0.659) |
| 50 | 0.943 (0.420) | 0.951 (0.539) | 0.971 (0.547) | 0.978 (0.625) |
| 100 | 0.921 (0.357) | 0.939 (0.448) | 0.965 (0.483) | 0.972 (0.559) |
| 500 | 0.914 (0.218) | 0.922 (0.258) | 0.961 (0.299) | 0.965 (0.346) |
| 1000 | 0.907 (0.173) | 0.911 (0.197) | 0.955 (0.237) | 0.959 (0.265) |
| 2000 | 0.900 (0.137) | 0.903 (0.153) | 0.951 (0.188) | 0.954 (0.205) |
| $\mu_1$ | | | | |
| $n$ | 90% CI | | 95% CI | |
| | T | E | T | E |
| 30 | 0.957 (0.544) | 0.849 (0.651) | 0.992 (0.624) | 0.899 (0.665) |
| 50 | 0.948 (0.539) | 0.876 (0.615) | 0.973 (0.620) | 0.908 (0.627) |
| 100 | 0.933 (0.519) | 0.883 (0.602) | 0.964 (0.617) | 0.917 (0.616) |
| 500 | 0.917 (0.415) | 0.889 (0.477) | 0.962 (0.548) | 0.934 (0.555) |
| 1000 | 0.907 (0.385) | 0.894 (0.424) | 0.957 (0.511) | 0.944 (0.525) |
| 2000 | 0.904 (0.350) | 0.899 (0.384) | 0.951 (0.471) | 0.948 (0.490) |
| $\mu_2$ | | | | |
| $n$ | 90% CI | | 95% CI | |
| | T | E | T | E |
| 30 | 0.960 (0.461) | 0.968 (0.620) | 0.992 (0.590) | 0.994 (0.672) |
| 50 | 0.949 (0.424) | 0.959 (0.541) | 0.977 (0.548) | 0.982 (0.630) |
| 100 | 0.925 (0.358) | 0.941 (0.472) | 0.970 (0.482) | 0.976 (0.539) |
| 500 | 0.915 (0.218) | 0.925 (0.304) | 0.961 (0.299) | 0.966 (0.348) |
| 1000 | 0.906 (0.173) | 0.914 (0.199) | 0.954 (0.237) | 0.958 (0.264) |
| 2000 | 0.901 (0.138) | 0.904(0.154) | 0.950 (0.188) | 0.954 (0.204) |
| $\mu_3$ | | | | |

Table 4.1: Coverage probabilities and length of the CI (in parentheses) using the true parameters (T) and the estimated parameters (E) for different sample sizes under $\mu_1$, $\mu_2$ and $\mu_3$.

Robinson (1997). As with (4.1), the dependence of $Y_i$'s and $\epsilon_i$'s on $n$ is suppressed but they must be viewed as triangular arrays. The extension to this setting would work along the following lines. The estimate of $\mu$ with dependent errors still has the same form as (4.2). With additional assumptions (Assumptions 1–5 of Robinson (1997)), it is guaranteed that $\sqrt{nh_n}(\hat{\mu}(x_i) - \mu(x_i))$, $x_i \in (0,1)$ and $x_1 \neq x_2$, converge jointly in distribution to independent normals with zero mean – a fact that justifies the consistency of our $p$-value based estimates in this setting. Hence, $\hat{d}_n$, defined using (4.3), can still be used to estimate the threshold. The limiting distribution would be of the same form as in Lemma 4.4 but with a different scaling factor that appears in the covariance function of the process $W$. We outline the form of the limiting distribution below. The technical details are more involved in the sense of tedium but the approach in deriving the limiting distribution remains the same at the conceptual level.

To precisely state the limiting distribution, let $\rho(i,j) = \rho(i-j)$ denote the covariance between $\epsilon_i$ and $\epsilon_j$ and let $\psi$ denote the underlying spectral density defined through the relation $\sigma_0^2\rho(l) = \int_{-\pi}^{\pi}\psi(u)\exp(\imath lu)du, l \in \mathbb{Z}$. Let $\bar{W}$ be a Gaussian process with drift $m(\cdot)$ (defined in Lemma 4.4) and covariance function

$$\mathrm{Cov}(\bar{W}(t_1), \bar{W}(t_2)) = 2\pi\psi(0)\int K(t_1+u)K(t_2+u)du.$$

It is not uncommon for the spectral density at zero, $\psi(0) = (2\pi)^{-1}\sigma_0^2\sum_{j\in\mathbb{Z}}\rho(j)$, to appear in settings with short range dependence (Robinson, 1997; Anevski and Hössjer, 2006).

**Proposition 4.13.** *Consider the setup of* (4.1) *with the errors now exhibiting short-range dependence as discussed above. Assume that for $h_n = h_0 n^{-1/(2k+1)}$, the resulting estimate $\hat{d}_n$ obtained using* (4.3) *satisfies $h_n^{-1}(\hat{d}_n - d_0) = O_p(1)$ and that the process*

$\bar{Z}(t) = \int_0^t [\Phi(\bar{W}(y)) - \gamma]dy$, $t \in \mathbb{R}$ *has a unique minimum a.s. Then*

$$h_n^{-1}(\hat{d}_n - d_0) = \underset{t\in\mathbb{R}}{\operatorname{argmin}}\, \bar{Z}(t).$$

The proof is outlined in Section C.7 of Appendix C.

An illustration of the above phenomenon is shown through a Q-Q plot (Figure 4.2), where we generate $\epsilon_i$'s from an AR(1) model $\epsilon_i = 0.25\epsilon_{i-1} + z_i$. Here, $z_i$s are mean 0 normal random variables with variance 0.0094 so that $\epsilon_i$'s have variance $(0.1)^2$. The Q-Q plot shows considerable agreement between the empirical quantiles, obtained from samples of size $n = 1000$, with the theoretical quantiles.

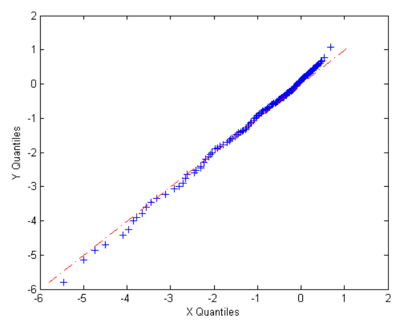



Figure 4.2: Regression setting with dependent errors: Q-Q plot under $\mu_1$.

## 4.6 Data Analysis

We now apply our procedure to two interesting examples from Chapter 2.

The first data set involves measuring concentration of mercury in the atmosphere through a LIDAR experiment which has a visible evidence of heteroscedasticity. The observed covariates can be considered to have arisen from a random design and the threshold $d_0$ corresponds to the distance at which there is a sudden rise in the concentration of mercury. We employ the non-normalized variant of our procedure (see Proposition 4.10) which is suited for heteroscedastic settings. It is reasonable to assume here that the function is at its baseline till range value 480. The estimate of $\tau_0$

is obtained by taking the average of observations until range reaches 480, which gives $\hat{\tau} = -0.0523$. The estimate $\hat{d}_n$ is obtained through the iterative approach described in Section 4.4. The expression for the approximate bias of the Nadaraya-Watson estimator turns out to be the same as that for the fixed design kernel estimator at $d_0$ while the approximate variance turns out to be $(\sigma_0^2 \bar{K^2})/(nh_n f(d_0))$ and the optimal value of $h_0$ is adjusted accordingly. The limiting distribution, as well as the optimal $h_0$, involves the parameter $\sigma_0(d_0)$, which we estimate using $\hat{\sigma}(\hat{d}_n)$ where $\hat{\sigma}^2(x) = \left[\mathbb{P}_n(Y - \hat{\mu}(X))^2 K\left((x - X)/h_n\right)\right] / \left[\mathbb{P}_n K\left((x - X)/h_n\right)\right]$.

The estimate $\hat{d}_n$ has an inherent bias which is a recurring feature in boundary estimation problems. A simple but effective way to reduce this bias is to subtract the median of the limiting distribution with imputed parameters, say $\hat{q}_{0.5}$, from our crude estimate, after proper normalization (so that the limiting median is zero). More precisely, $\hat{d}_n - n^{-1/(2k+1)}\hat{q}_{0.5}$ is our final estimate. Assuming $k$ to be 1, the resulting estimate of $d_0$ is 551.05 which appears reasonable (see Figure 2.1 of Chapter 2). Moreover, the CIs are $[550.53, 555.17]$ and $[549.75, 557.82]$ for confidence levels of 90% and 95%, respectively, which also seem reasonable.

Our second data set, which comes from the last example in Chapter 2, involves the measurement of annual global temperature anomalies, in degree Celsius, over the years 1850 to 2009. The depiction of the data (see Figure 2.1 of Chapter 2) suggests a trend function which stays at its baseline value for a while followed by a nondecreasing trend. We follow the approach of Wu et al. (2001) and Zhao and Woodroofe (2012), and model the data as having a non-parametric trend function and short-range dependent errors. The flat stretch at the beginning is also noted in Zhao and Woodroofe (2012), where isotonic estimation procedures are considered in settings with dependent data. They also provide evidence for the errors to be arising from a lower order auto-regressive process. A comprehensive approach would incorporate a cyclical component as well (Schlesinger and Ramankutty, 1994), which we do not pursue here.

The estimate of the baseline value, after averaging the anomalies up to the year 1875, is $\hat{\tau} = -0.3540$. Using this estimate of $\tau_0$, we employ our procedure with non-normalized $p$-values (see (4.3)) in this example with the optimal $h_0$ chosen through an iterative approach. Constructing the CI involves estimating an extra parameter $\psi(0)$ for which we use the estimates computed in Wu et al. (2001, pp. 800) (the parameter $\sigma^2$ estimated in that paper is precisely $2\pi\psi(0)$). Assuming $k$ to be 1, the estimate of the threshold $d_0$ after bias correction, which signifies the advent of global warming, turns out to be 1912. The CIs are $[1908, 1917]$ and $[1906, 1919]$ for confidence levels 90% and 95% respectively. This is compatible with the observation on page 2 of Zhao and Woodroofe (2012) that global warming does not appear to have begun until 1915.

## 4.7 Conclusion

*Adaptivity.* As was the case with the dose-response setting, we can come with an adaptive approach that yields (one-sided) upper confidence intervals for $d_0$ when $k$ is unknown. Also, when $k$ is unknown, ideas from multiscale testing procedures for white noise models (Dümbgen and Spokoiny, 2001; Dümbgen and Walther, 2008) can conceivably be used to develop adaptive procedures in our model. This is a hard open problem and a topic for future research.

*Minimaxity.* The estimators studied in this chapter attain the convergence rate of $n^{-1/(2k+1)}$. This leads to a natural question as to whether this is the best possible rate of convergence. When $\mu$ is monotone increasing, $d_0$ is precisely $\mu^{-1}(\tau_0)$, where $\mu^{-1}$ is the right continuous inverse of $\mu$. Wright (1981) (Theorem 1) shows that the rate of convergence of the isotonic least squares estimate $\mu$ at a point, $x_0$, where the first $k-1$ derivatives vanish but the $k$th does not, is precisely $n^{-k/(2k+1)}$. A slightly more general result establishing a process convergence is stated in Fact 1 of Banerjee (2009). Using this in conjunction with the techniques for the proof of Theorem 1 in Banerjee and Wellner (2005), it can be deduced that the rate of convergence of the isotonic estimate of $\mu^{-1}$ at $\mu(x_0)$ is $n^{-1/(2k+1)}$, which matches the rate attained by our approach. Hence, we expect this rate to be minimax in our setting. We note that this rate is not the same as the faster rate $\min(n^{-2/(2k+3)}, n^{-1/(2k+1)})$ obtained in Neumann (1997) for a change-point estimation problem in a density deconvolution model and also observed in the convolution white noise models of Goldenshluger et al. (2006) and Goldenshluger et al. (2008). These models are related to our setting; e.g., Problem 1 in Goldenshluger et al. (2008) is a Gaussian white noise model where the underlying regression function also has a cusp of a known order at an unknown point of interest. The convolution white noise model considered in Goldenshluger et al. (2006) (Problem 2 in Goldenshluger et al. (2008)) is equivalent to this problem for a particular choice of the convolution operator; see Goldenshluger et al. (2006, pp. 352–353) and Goldenshluger et al. (2008, pp. 790–791) for more details. Besides these being white noise models, they differ from our setting through an additional smoothness condition (Goldenshluger et al., 2006, pp. 354–355), which translates, in our setting, to assuming that $\mu^{(k)}$ is Lipschitz outside any neighborhood of $d_0$, an assumption *not made* in this chapter. Hence, Neumann's rate need not be minimax for our setting. The faster rate of Neumann (1997) was also observed for $k = 1$ in Cheng and Raimondo (2008) but once again under the assumption that the derivative of the regression function is at least twice differentiable away from the change-point, again an assumption *not made* for this approach.

# CHAPTER 5

# Baseline zone estimation in two-dimensions

In this Chapter, we address the two-dimensional version of the threshold estimation problem from Chapter 2. In particular, we consider a model of the form $Y = \mu(X) + \epsilon$, where $\mu$ is a function on $[0,1]^2$ such that

$$\mu(x) = \tau_0 \text{ for } x \in S_0, \text{ and } \mu(x) > \tau_0 \text{ for } x \notin S_0 \tag{5.1}$$

and $\tau_0$ is unknown. The covariate $X$ may arise from a random or a fixed design setting and we assume that $\epsilon$ has mean zero with finite positive variance $\sigma_0^2$. Interest centers on estimating the baseline region $S_0$ beyond which the function deviates from its baseline value. There are several practical motivations behind detecting $S_0$ (or $S_0^c$) which can be thought of as the region of no-signal. For example, in several *fMRI* studies, one seeks to detect regions of brain activity from cross sectional two-dimensional images. Here, $S_0$ corresponds to the region of no-activity in the brain with $S_0^c$ being the region of interest. In the two dimensional version of *LIDAR (light detection and ranging)* experiments used for measuring concentration of pollutants in the atmosphere, interest often centers on finding high/low pollution zones (see, for example, Wakimoto and McElroy (1986)); in such contexts, $S_0$ would be the zone of minimal pollution. In *dose-response studies*, patients may be put on multiple (interacting) drugs (see, for example, Geppetti and Benemei (2009)), and it is of interest to find the dosage levels ($\partial S_0$) at which the drugs starts being effective.

The question of detecting $S_0$ is also related to the *edge detection* problem which involves recovering the boundary of an image. In edge detection, $\mu$ corresponds to the image intensity function with $S_0^c$ being the image and $S_0$ the background. A number of different algorithms in the computer science literature deal with this problem, though primarily in situations where $\mu$ has a jump discontinuity at the boundary of $S_0$; see Qiu (2007) for a review of edge detection techniques. With the exception of work done by Korostelëv and Tsybakov (1993b), Mammen and Tsybakov (1995) and a few others, theoretical properties of such algorithms appear to have been rarely addressed. In fact, the study of theoretical properties of such estimates is typically intractable without some regularity assumption on $S_0$; for example, Mammen and Tsybakov (1995) discuss minimax recovery of sets under smoothness assumption on the boundary.

In this chapter, we approach the problem from the point of view of a shape-constraint (typically obtained from background knowledge) on the baseline region. We assume that the region $S_0$ is a closed convex subset of $[0, 1]^2$ with a non-empty interior (and therefore, positive Lebesgue measure) and restrict ourselves to the more difficult problem where $\mu$ is continuous at the boundary. Convexity is a natural shape restriction to impose, not only because of analytical tractability, but also as convex boundaries arise naturally in several application areas: see, Wang et al. (2007), Ma et al. (2010), Stahl and Wang (2005) and Goldenshluger and Spokoiny (2006) for a few illustrative examples. In the statistics literature, Goldenshluger and Zeevi (2006) provide theoretical analyses of a convex boundary recovery method in a white noise framework. While this has natural connections to our problem, we note that they impose certain conditions (see Definitions 2 and 3 of Goldenshluger and Zeevi (2006) and the associated discussions), which restricts the geometry of the set of interest, $G$, beyond convexity. Hence, their results, particularly on the rate of convergence, are difficult to compare to the ones obtained in our problem. Further, they estimate $G$ through its *support function* which needs to be estimated along all directions. It is unclear whether an effective algorithm can be devised to adopt this procedure in a regression setting.

Our problem also has connections to the level-sets estimation problem since $S_0^c$ is the "level-set" $\{x : \mu(x) > \tau_0\}$ of the function $\mu$. However, because $\tau_0$ is at the extremity of the range of $\mu$, the typical level-set estimate $\{x : \hat{\mu}(x) > \tau_0\}$, where $\hat{\mu}$ is an estimate of $\mu$, does not perform well unless $\mu$ has a jump at $\partial S_0$ (a situation not considered in this chapter). Moreover, this plug-in approach does not account for the pre-specified shape of the level-set. We note that the shape-constrained approach to estimate level-sets has received some attention in the literature, e.g., Nolan (1991) studied estimating ellipsoidal level-sets in the context of densities, Hartigan (1987) provided an algorithm for estimating convex contours of a density, and Tsybakov (1997) and Cavalier (1997) studied "star-shaped" level-sets of density and regression functions respectively. All the above approaches are based on an "excess mass" criterion (or its local version) that yield estimates with optimal convergence rates (Tsybakov, 1997). It will be seen later that our estimate also recovers the level-set of a transform of $\mu$, but at a level in the *interior* of the range of the transform. More connections in this regard are explored in Section 5.4.

In this chapter, we extend the $p$-value based approach from Chapter 2. The extension is fairly involved, even from a computational perspective (see Section 5.1.1). The smoothness of $\mu$ at its boundary again plays a critical role in determining the rate of convergence of our estimate: for a regression function which is "$p$-regular" (formally defined in Section 5.2) at the boundary of the convex baseline region, our estimate converges at a rate $N^{-2/(4p+3)}$ in the dose–response setting, $N$ being the total budget. This coincides with the minimax rate of a related level-set estimation problem; see Polonik (1995, Theorem 3.7) and Tsybakov (1997, Theorem 2). The analogue of the estimate in the regression setting converges at the slightly slower rate of $N^{-1/(2p+2)}$. The difference in the two rates is due to the bias introduced from the use of kernel estimates in the regression setting. A more technical explanation is given in Remark 5.15. It should be pointed out that our convergence rates are very different from the analogous problem in the density estimation scenario which corresponds to finding the support of a multivariate density. Faster convergence rates (Härdle et al., 1995) can be obtained in density estimation due to the simpler nature of the problem: namely, there are *no realizations* from outside the support of the density.

The main contributions of this chapter are the following. We extend a novel and computationally simple $p$-value approach to estimate baseline sets in two dimensions and deduce consistency and rates of convergence of our estimate in the two aforementioned settings. Our approach falls at the interface of edge detection and level-set estimation problem as it detects the edge set ($S_0^c$) through a level-set estimate (see Section 5.4). The proofs require heavy-duty applications of non-standard empirical processes and, along the way, we deduce results which may be of independent interest. For example, we apply a blocking argument which leads to a version of Hoeffding's inequality for $m$-dependent random fields, which is then further extended to an empirical process inequality. This should find usage in spatial statistics and is potentially relevant to approaches based on $m$-approximations that answer the central limit question for dependent random fields and their empirical process extensions; see Rosén (1969), Bolthausen (1982) and Wang and Woodroofe (2013) for some work on $m$-dependent random fields and $m$-approximations. While we primarily address the situation where the baseline set is convex, in the presence of efficient algorithms, our approach is extendible beyond convexity (see Section 5.4).

The rest of the chapter is arranged as follows: we briefly describe the estimation procedure in the two settings in Section 5.1. Barring $\mu$ and $S_0$, notations are *not carried forward* from the dose-response setting to the regression setting unless stated otherwise. We list our assumptions in Section 5.2. We justify consistency and deduce an upper bound on the rate of the convergence of our procedure (assuming a known $\tau_0$) for the dose-response and regression settings in Sections 5.3.1 and 5.3.2 respectively. Situations with an unknown $\tau_0$ are addressed in Section 5.3.3. We explore extensions to non-convex baseline regions and connections with level-set estimation in Section 5.4.

## 5.1 Estimation Procedure

In this section, we extend the non-normalized variant of the $p$-value procedure from Chapter 2. The two versions of the procedure, the normalized and the non-normalized one, exhibit similar fundamental features such as the same dichotomous separation over $S_0$ and $S_0^c$, and identical rates of convergence. The non-normalized version is notationally more tractable and avoids a few routine justifications required for the normalized version.

### 5.1.1 Dose-Response Setting

Consider a model of the form

$$Y_{ij} = \mu(X_i) + \epsilon_{ij}, \quad j = 1, \ldots, m, \quad i = 1, \ldots, n.$$

Here $m = m_n = m_0 n^\beta$ for some $\beta > 0$, with $N = m \times n$ being the total budget. The covariate $X$ is sampled from a distribution $F$ with Lebesgue density $f$ on $[0,1]^2$ and $\epsilon$ is independent of $X$, has mean 0 and variance $\sigma_0^2$.

Let $\bar{Y}_{i\cdot} = \sum_{j=1}^m Y_{ij}/m$ and $\hat{\tau}$ is some suitable estimate of $\mu$ (to be discussed later). Following the approach from Chapter 2, an estimate of $S_0$ can be constructed by minimizing

$$\mathbb{M}_n(S) = \mathbb{P}_n \left\{ \Phi\left(\sqrt{m}(\bar{Y} - \hat{\tau})\right) - \gamma \right\} 1_S(X),$$

where $\mathbb{P}_n$ denotes the empirical measure on $\{\bar{Y}_{i\cdot}, X_i\}_{i \le n}$ and $\gamma = 3/4$.

The class of sets over which $\mathbb{M}_n$ is minimized should be chosen carefully as very large classes would give uninteresting discrete sets while small classes may not provide a reasonable estimate of $S_0$. As we assumed $S_0$ to be convex, we minimize $\mathbb{M}_n$ over $\mathcal{S}$, the class of *closed convex* subsets of $[0,1]^2$. Let $\hat{S}_n = \mathrm{argmin}_{S \in \mathcal{S}} \mathbb{M}_n(S)$. The estimate $\hat{S}_n$ can be computed by an adaptation of a density level-set estimation algorithm (Hartigan, 1987) which we state below. Note that if a closed convex set $S^\star$ minimizes $\mathbb{M}_n$, the convex hull of $\{X_i : X_i \in S^\star, 1 \le i \le n\}$ also minimizes $\mathbb{M}_n$. Hence, it suffices to reduce our search to convex polygons whose vertices belong to the set of $X_i$'s. There could be $2^n$ such polygons. So, an exhaustive search is computationally expensive.

*Computing the estimate.* We first find the optimal polygon (the convex polygon which minimizes $\mathbb{M}_n$) for each choice of $X$ as its leftmost vertex. We use the following notation. Let this particular $X$ be numbered 1, and let the $X_i$'s not to its left be numbered $2, 3, ..., r$. The axes are shifted so that 1 is at the origin and the coordinates of point $i$ are denoted by $z_i$. The line segment $az_i + (1-a)z_j$, $(0 \le a \le 1)$ is written as $[i, j]$. Assume that $1, ..., r$ are ordered so that the segments $[1, i]$ move counterclockwise as $i$ increases and so that $i \le j$ if $i \in [1, j]$. Polygons will be built up from triangles for $1 < i < j \le r$; $\Delta_{ij}$ is the convex hull of $(1, i, j)$ excluding $[1, i]$. Note that the segment $[1, i]$ is excluded from $\Delta_{ij}$ in order to combine triangles without overlap. The

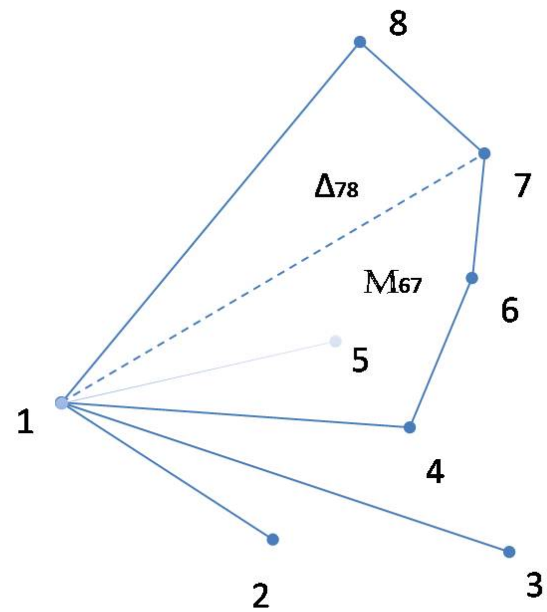



Figure 5.1: Notation for constructing the convex set estimate. An arbitrary vertex is numbered 1, and those not to its left are numbered $2, 3, \ldots, 8$ in a counterclockwise manner. The triangle $\Delta_{78}$ excludes the line segment $[1, 7]$. The optimal polygon (with measure $\mathbb{M}_{67}$) with successive vertices 6, 7 and 1 is depicted as the convex polygon with vertices 1,4,6 and 7.

quadrilateral with vertices at $1, i, j, k$ for $i < j < k \leq r$ is convex if

$$D_{ijk} = \begin{vmatrix} z_i' & 1 \\ z_j' & 1 \\ z_k' & 1 \end{vmatrix} \geq 0$$

Let $\mathbb{M}_{1j}$ be the value of $\mathbb{M}_n$ on the line segment $[1, j]$. Further, for $1 < j < k \leq r$, let $\mathbb{M}_{jk}$ denote the minimum value of $\mathbb{M}_n$ among closed convex polygons with successive counterclockwise vertices $j, k$ and 1. Note that all such convex polygons contain the triangle $\Delta_{jk}$ and hence, $\mathbb{M}_n(\Delta_{jk})$, $\mathbb{M}_n$ measure of $\Delta_{jk}$, is a common contributing term to the $\mathbb{M}_n$ measure of all such polygons. This simple fact forms the basis of the algorithm. It can be shown that

$$\mathbb{M}_{jk} = \mathbb{M}_{i^*j} + \mathbb{M}_n(\Delta_{jk}), \tag{5.2}$$

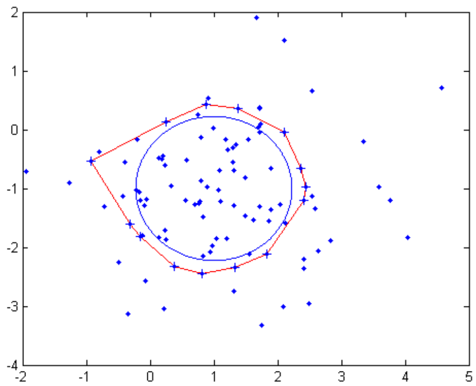



Figure 5.2: An illustration of the procedure in the dose-response setting with $m = 10$ and $n = 100$. The set $S_0$ is a circle centered at $(1, -1)$ with radius 1.

where $i^* = I(k, j)$ is chosen to minimize $\mathbb{M}_{ij}$ over vertices $i$ with $i < j$, $D_{ijk} \geq 0$, i.e,

$$i^* = I(k, j) = \underset{i:i<j, D_{ijk}>0}{\operatorname{argmin}} \mathbb{M}_{ij}. \tag{5.3}$$

Note that $i^*$ could possibly be 1, in which case $\mathbb{M}_{jk}$ is simply the $\mathbb{M}_n$ measure of the triangle formed by $j$, $k$ and 1 (including the contribution of line segment $[1, j]$).

One way to construct an optimal polygon with leftmost vertex 1 is to find the minimum among $\mathbb{M}_{jk}$, $1 \leq j < k$, where $\mathbb{M}_{jk}$'s are computed recursively using 5.2 and 5.3. Hence, one optimal polygon with leftmost vertex 1 has vertices $i_l, i_2, \ldots, i_s = 1$, where either $s = 1$ or $\mathbb{M}_{i_2 i_1} = \min_{1 \leq j < k} \mathbb{M}_{jk}$, $i_3 = I(i_1, i_2)$, $i_4 = I(i_2, i_3)$, $\ldots$, $1 = i_s = I(i_{s-2}, i_{s-1})$. Once this is done for each choice of $X$ as the leftmost vertex, the final estimate $\hat{S}_n$ is simply the one with the minimum $\mathbb{M}_n$ value among these $n$ constructed polygons.

There are minor modifications to this algorithm which reduce the over-all implementation to $O(n^3)$ computations; see Hartigan (1987, Section 3) for more details.

### 5.1.2 Regression Setting

Consider a model of the form

$$Y_{kl} = \mu(x_{kl}) + \epsilon_{kl},$$

with $x_{kl} = (u_k, v_l)$, $u_k = k/m$, $v_l = l/m$, $k, l \in \{1, \ldots, m\}$. The total number of observations is thus $n = m^2$. The errors $\epsilon_{kl}$s are independent with mean 0 and variance $\sigma_0^2$. Here, $\mu$ is as defined earlier and we seek to estimate $S_0 = \mu^{-1}(0)$.

Let

$$\hat{\mu}(x) = \frac{1}{nh_n^2} \sum_{k,l} Y_{kl} K\left(\frac{x - x_{kl}}{h_n}\right)$$

denote the estimator of $\mu$, with $K$ being a probability density (kernel) on $\mathbb{R}^2$ and $h_n$ the smoothing bandwidth. We take $h_n = h_0 n^{-\beta}$ for $\beta < 1/2$ and $K$ to be the 2-fold product of a symmetric one-dimensional compact kernel, i.e., $K(x_1, x_2) = K_0(x_1)K_0(x_2)$, where $K_0$ is a symmetric probability density on $\mathbb{R}$ with $K_0(x) = 0$ for $|x| \geq L_0$. Note that (the normalized) $\hat{\mu}$ is asymptotically normal and the limiting variance of $\sqrt{nh_n^2}\hat{\mu}(x)$ is $\Sigma^2 = \sigma_0^2 \int_{u \in \mathbb{R}^2} K^2(u)du > 0$ and hence, the multiple $\sqrt{nh_n^2}$ is used (instead of $\sqrt{nh_n}$) to construct the $p$-values. Let $\hat{\tau}$ is a suitable estimate of $\tau_0$. We can estimate $S_0$ by minimizing

$$\begin{aligned} \mathbb{M}_n(S) &= \frac{1}{n} \sum_{k,l: x_{kl} \in \mathcal{I}_n} \left\{ \Phi\left(\sqrt{nh_n^2}(\hat{\mu}(x_{kl}) - \hat{\tau})\right) - \frac{3}{4} \right\} 1_S(x_{kl}) \\ &= \frac{1}{n} \sum_{k,l: x_{kl} \in \mathcal{I}_n} \widetilde{W}_{kl} 1_S(x_{kl}) \end{aligned} \tag{5.4}$$

with $\widetilde{W}_{kl} = \Phi\left(\sqrt{nh_n^2}\hat{\mu}(x_{kl})\right) - \gamma$ and $\gamma = 3/4$. To avoid the bad behavior of the kernel estimator at the boundary, the sums are restricted to design points in $\mathcal{I}_n = [L_0 h_n, 1 - L_0 h_n]^2$. With $\mathcal{S}$ being the class of *closed convex* subsets of $[0, 1]^2$ as defined earlier, let $\hat{S}_n = \text{argmin}_{S \in \mathcal{S}} \mathbb{M}_n(S)$.

The estimate can be computed using the algorithm stated at the end of Section 5.1.1

## 5.2 Notations and Assumptions

We adhere to the setup of Sections 5.1.1 and 5.1.2, i.e., we assume the errors to be independent and homoscedastic and consider random and fixed designs respectively for the dose-response and regression settings. A fixed design in the regression setting provides a simpler platform to illustrate the main techniques. In particular, it allows to treat the kernel estimates as an $m'$–dependent random field which facilitates in obtaining probability bounds on our estimate; see Section 5.3.2. Also, a random design in the dose–response setting permits the use of empirical process techniques developed for i.i.d. data ($(\bar{Y}_i, X_i)$'s are i.i.d.). However, we note here that the dose-response model in a fixed (uniform) design setting can be addressed by taking an approach similar (and in fact, simpler due to the absence of smoothing) to that for the regression setting. The results on the rate of convergence of our estimate of $S_0$ are identical for the random design and the fixed uniform design dose-response models.

Let $\lambda$ denote the Lebesgue measure. The precision of the estimates is measured using the metrics

$$d_F(S_1, S_2) = F(S_1 \Delta S_2) \text{ and } \; d(S_1, S_2) = \lambda(S_1 \Delta S_2)$$

for the dose–response and the regression settings respectively. The two metrics arise naturally in their respective settings as $X_i$'s have distribution $F$ (in the dose–response setting) and the empirical distribution of the grid points in the regression setting converges to the Uniform distribution on $[0, 1]^2$.

For simplicity, we assume $\tau_0$ to be known. It can be shown that our results would extend to case where we impute a $\sqrt{mn}$ (dose-response)/ $\sqrt{n}$ (regression) estimate of

$\tau$ (more on this in Section 5.3.3). We summarize the assumptions below:

1. The function $\mu$ is continuous on $[0,1]^2$. For the standard regression setting, we additionally assume that $\mu$ is Lipschitz continuous of order 1 .

2. The function $\mu$ is $p$-regular at $\partial S_0$, i.e., for some $\kappa_0, C_0, C_1 > 0$ and for all $x \notin S_0$ such that $\rho(x, S_0) < \kappa_0$,

$$C_0 \rho(x, S_0)^p \leq \mu(x) - \tau_0 \tag{5.5}$$

   Here $\rho$ is the $\ell_\infty$ metric in $\mathbb{R}^2$ (for convenience).

3. $S_0 = \mu^{-1}(\tau_0)$ is convex. For some $\epsilon_0 > 0$, , $S_0 \subset [\epsilon_0, 1-\epsilon_0]^2$ and $\lambda(S_0) > 0$.

4. The design density $f$ for the dose-response setting is assumed to be continuous and positive on $[0,1]^2$.

5. Assumptions on the kernel $K(x) = K_0(x_1)K_0(x_2)$, $x = (x_1, x_2)$, for the standard regression setting:

   (a) $K_0$ is a symmetric probability density.

   (b) $K_0$ is compactly supported, i.e., $K_0(x) = 0$ when $|x| \geq L_0$, for some $L_0 > 0$.

   (c) $K$ is Lipschitz continuous of order 1.

Note that by uniform continuity of $\mu$ and compactness of $[0,1]^2$, $\inf\{\mu(x) : d(x, S_0) \geq \kappa_0\} > \tau_0$. For a fixed $p$, $\tau_0$, $\kappa_0$, $\delta_0 > 0$, we denote the class of functions $\mu$ satisfying assumptions 1, 2, 3 and

$$\inf\{\mu(x) : d(x, S_0) \geq \kappa_0\} - \tau_0 > \delta_0 \tag{5.6}$$

by $\mathcal{F}_p = \mathcal{F}_p(p, \tau_0, \kappa_0, \delta_0)$.

**Remark 5.1.** *It can be readily seen that if the regularity assumption in* (5.5) *holds for a particular* $p$*, it also holds for any* $\tilde{p} > p$ *as well. We assume that we are working with the smallest* $p$ *such that* (5.5) *is satisfied (the set of values* $\tilde{p}$ *such that* (5.5) *holds for a fixed* $\mu$*,* $C_0$ *and* $\kappa_0$ *is a closed set and is bounded from below whenever it is non-empty). In level-sets estimation theory, analogous two-sided conditions of the form*

$$C_0\rho(x, S_0)^p \leq |\mu(x) - \tau_0| < C_1\rho(x, S_0)^p$$

*are typically assumed (see Tsybakov (1997, Assumptions (4) and (4')), Cavalier (1997, Assumption (4))). This stronger condition restricts the choice of* $p$*. However, we note here that the left inequality plays a more significant role as it provides a lower bound on the amount by which* $\mu(x)$ *differs from* $\tau_0$ *in the vicinity of* $\partial S_0$*. Some results in a density level-set estimation problem with a slightly weaker analogue of the left inequality can be found in Polonik (1995). The upper bound (right inequality) is seen to be useful for establishing adaptive properties of certain density level-set estimates (Singh et al., 2009).*

## 5.3 Consistency and Rate of Convergence

### 5.3.1 Dose-response setting

As $\tau_0$ is known, we take $\tau_0 = 0$ without loss of generality. Recall that $\mathbb{M}_n(S) = \mathbb{P}_n \left\{\Phi\left(\sqrt{m}\bar{Y}\right) - \gamma\right\} 1_S(X)$. Let $P_m$ denote the measure induced by $(\bar{Y}, X)$ and

$$M_m(S) = P_m\left[\left\{\Phi\left(\sqrt{m}\bar{Y}\right) - \gamma\right\} 1_S(X)\right].$$

The process $M_m$ acts as a population criterion function and can be simplified as follows. Let

$$Z_{1m} = \frac{1}{\sqrt{m}\sigma_0}\sum_{j=1}^{m} \epsilon_{1j} \tag{5.7}$$

and $Z_0$ be a standard normal random variable independent of $Z_{1m}$s. Then

$$\begin{aligned} E\left[\Phi\left(\sqrt{m}\bar{Y}_1\right)\big|\, X_1 = x\right] &= E\left[\Phi\left(\sqrt{m}\mu(x) + \sigma_0 Z_{1m}\right)\right] \\ &= E\left[E\left[1\left(Z_0 < \sqrt{m}\mu(x) + \sigma_0 Z_{1m}\right)\big|\, Z_{1m}\right]\right] \\ &= P\left[\frac{Z_0 - \sigma_0 Z_{1m}}{\sqrt{1+\sigma_0^2}} < \frac{\sqrt{m}\mu(x)}{\sqrt{1+\sigma_0^2}}\right] = \Phi_m\left(\frac{\sqrt{m}\mu(x)}{\sqrt{1+\sigma_0^2}}\right), \end{aligned}$$

where $\Phi_m$ denotes the distribution function of $(Z_0 - \sigma_0 Z_{1m})/\sqrt{1+\sigma_0^2}$. By Pólya's theorem, $\Phi_m$ converges uniformly to $\Phi$ as $m \to \infty$. Hence, it can be seen that

$$\lim_{m\to\infty} E\left[\Phi\left(\sqrt{m}\bar{Y}_1\right)\big|\, X_1 = x\right] = \frac{1}{2}1_{S_0}(x) + 1_{S_0^c}(x).$$

By the Dominated Convergence Theorem, $M_m(S)$ converges to $M(S)$, where

$$\begin{aligned} M(S) = M_F(S) &= \int_S \left(\frac{1}{2}1_{S_0}(x) + 1_{S_0^c}(x) - \gamma\right) F(dx) \\ &= (1/2 - \gamma)F(S_0 \cap S) + (1-\gamma)F(S_0^c \cap S). \end{aligned} \tag{5.8}$$

Note that $S_0$ minimizes the limiting criterion function $M(S)$. An application of the argmin continuous mapping theorem (van der Vaart and Wellner, 1996, Theorem 3.2.2) yields the following result on the consistency of $\hat{S}_n$

**Theorem 5.2.** *Assume $S_0$ to be a closed convex set and the unique minimizer of $M(S)$. Then $\sup_{S\in\mathcal{S}} |\mathbb{M}_n(S) - M(S)|$ and $d_F(\hat{S}_n, S_0)$ converge in outer probability to zero for any $\gamma \in (0.5, 1)$.*

**Remark 5.3.** *We end up proving a stronger result. The consistency is established in terms of the Hausdorff metric which implies consistency with respect to $d_F$. Moreover, we do not require $m$ to grow as $m_0 n^\beta$, $\beta > 0$ for consistency. The condition $\min(m, n) \to \infty$ suffices. Also, the result extends to higher dimensions as well, i.e., when $\mu$ is a function from $[0,1]^d \mapsto \mathbb{R}$ and $S_0 = \mu^{-1}(0)$ is a closed convex subset of*

$[0,1]^d$, *then the analogous estimate is consistent. However an efficient way to compute the estimate is not immediate.*

The proof is given in Section D.1 of Appendix D.

We now proceed to deducing the rate of convergence of $d_F(\hat{S}_n, S_0)$. For this, we study how small the difference $(\mathbb{M}_n - M)$ is and how $M$ behaves in the vicinity of $S_0$. We split the difference $(\mathbb{M}_n - M)$ into $(\mathbb{M}_n - M_m)$ and $(M_m - M)$ and study them separately. The term $\mathbb{M}_n - M_m$ involves an empirical average of centered random variables, efficient bounds on which are derived using empirical process inequalities. We start with establishing a bound on the non-random term $(M_m - M)$ in the vicinity of $S_0$. To this end, we first state a fact that gets frequently used in the proofs that follow. For any $\delta > 0$, let $S^\delta = \{x : \rho(x, S) < \delta\}$ and ${}_\delta S = \{x : \rho(x, S_0^c) \geq \delta\}$ denote the $\delta$-fattening and $\delta$-thinning of the set $S$. There exists a constant $c_0 > 0$ such that for any $S \in \mathcal{S}$,

$$\lambda(S^\delta \backslash {}_\delta S) \leq c_0 \delta \text{ and consequently, } F(S^\delta \backslash {}_\delta S) \leq \tilde{c}_0 \delta, \tag{5.9}$$

with $\tilde{c}_0 = \|f\|_\infty c_0$ ($\|f\|_\infty < \infty$, by Assumption 4). For a proof of the above, see, for example, Dudley (1984, pp. 62–63). We now address the non-random term $(M_m - M)$.

**Lemma 5.4.** *For any $\delta > 0$, $a_n \downarrow 0$ and $S \in \mathcal{S}$ such that $F(S \triangle S_0) < \delta$,*

$$\begin{aligned} |(M_m - M)(S) - (M_m - M)(S_0)| \;\leq\; & |\Phi_m(0) - 1/2|\delta + \min(\tilde{c}_0 a_n, \delta) \\ & + \left| \Phi_m\left( \frac{C_0 \sqrt{m} a_n^p}{\sqrt{1+\sigma_0^2}} \right) - 1 \right| \delta \\ & + \left| \Phi_m\left( \frac{\sqrt{m}\delta_0}{\sqrt{1+\sigma_0^2}} \right) - 1 \right| \delta. \end{aligned}$$

*Proof.* Note that

$$\begin{aligned} M_m(S) - M_m(S_0) &= P_m\left[\left\{\Phi_m\left(\frac{\sqrt{m}\mu(x)}{\sqrt{1+\sigma_0^2}}\right) - \gamma\right\}\{1_S(x) - 1_{S_0}(x)\}\right] \text{ and} \\ M(S) - M(S_0) &= \int \{(1/2)1_{S_0}(x) + 1_{S_0^c}(x) - \gamma\}\{1_S(x) - 1_{S_0}(x)\} F(dx). \end{aligned}$$

Hence, the expression $|(M_m - M)(S) - (M_m - M)(S_0)|$ is bounded by

$$\int_{x\in(S_0\cap S)} \left|\Phi_m(0) - \frac{1}{2}\right| F(dx) + \int_{x\in(S_0^c\cap S)} \left|\Phi_m\left(\frac{\sqrt{m}\mu(x)}{\sqrt{1+\sigma_0^2}}\right) - 1\right| F(dx). \quad (5.10)$$

Note that the first term is bounded by $|\Phi_m(0) - 1/2|\delta$. Further, let $S_n = \{x : \rho(x, S_0) \geq a_n\}$. Using (5.9), $F(S_n^c \backslash S_0) \leq \tilde{c}_0 a_n$. Also, as $a_n \downarrow 0$, $a_n < \kappa_0$ for sufficiently large $n$. Thus, for $x \in S_n$,

$$\mu(x) \geq \min(\rho(x, S_0)^p, \delta_0) \geq \min(a_n^p, \delta_0),$$

using (5.5) and (5.6). Hence, the second sum in (5.10) is bounded by

$$\begin{aligned} F(S_n^c\backslash S_0) + \int_{x\in(S_n\cap S)} \left|\Phi_m\left(\frac{\sqrt{m}\mu(x)}{\sqrt{1+\sigma_0^2}}\right) - 1\right| F(dx) \leq \min(\tilde{c}_0 a_n, \delta) \\ + \int_{x\in(S_n\cap S)} \left\{\left|\Phi_m\left(\frac{C_0\sqrt{m}a_n^p}{\sqrt{1+\sigma_0^2}}\right) - 1\right| + \left|\Phi_m\left(\frac{\sqrt{m}\delta_0}{\sqrt{1+\sigma_0^2}}\right) - 1\right|\right\} F(dx). \end{aligned}$$

As $F(S_n \cap S) < \delta$, we get the result. □

To control $\mathbb{M}_n - M_m$, we rely on a version of Theorem 5.11 of van de Geer (2000). The result in its original form is slightly general. In their notation, it involves a bound on a special metric $\rho_K(\cdot)$ (see van de Geer (2000, equation 5.23)) which, in light of Lemma 5.8 of van de Geer (2000), can be controlled by bounding the $L_2$-norm in the case of bounded random variables. This yields the consequence stated below. Here,

$H_B$ denotes the entropy with respect to bracketing numbers.

**Theorem 5.5.** *Let $\mathcal{G}$ be a class of functions such that $\sup_{g\in\mathcal{G}} \|g\|_\infty \le 1$. For some universal constant $C > 0$, let $C_2, C_3, R$ and $N > 0$ satisfy the following conditions:*

$$
\begin{aligned}
R &\ge \sup_{g\in\mathcal{G}} \|g\|_{L_2(P)}, \\
N &\ge C_2 \int_0^R H_B^{1/2}(u, \mathcal{G}, L_2(P))du \vee R \\
C_2^2 &\ge C^2(C_3+1) \ \textit{and} \\
N &\le C_3\sqrt{n}R^2.
\end{aligned}
$$

*Then*

$$P^*\left[\sup_{g\in\mathcal{G}} |\mathbb{G}_n(g)| > N\right] \le C\exp\left[\frac{-N^2}{C^2(C_3+1)R^2}\right],$$

*where $P^*$ denotes the outer probability.*

We have the following proposition on the rate of convergence of $\hat{S}_n$.

**Proposition 5.6.** *When $\beta > 0$,*

$$P^*\left(d_F(\hat{S}_n, S_0) > \delta_n\right) \to 0$$

*for $\delta_n = K_1 \max\{n^{-2/3}, m^{-1/(2p)}\}$, where $K_1 > 0$ is some constant.*

*Proof.* Let $k_n$ be the smallest integer such that $2^{k_n+1}\delta_n \ge 1$. For $0 \le k \le k_n$, let $\mathcal{S}_{n,k} = \left\{S : S \in \mathcal{S},\ 2^k\delta_n < d_F(S, S_0) \le 2^{k+1}\delta_n\right\}$. As $\hat{S}_n$ is the minimizer for $\mathbb{M}_n$,

$$P^*\left(d_F(\hat{S}_n, S_0) > \delta_n\right) \le \sum_{k=0}^{k_n} P^*\left(\inf_{A\in\mathcal{S}_{n,k}} \mathbb{M}_n(A) - \mathbb{M}_n(S_0) \le 0\right).$$

The sum on the right side is bounded by

$$\sum_{k=0}^{k_n} P^* \left( \sup_{A \in \mathcal{S}_{n,k}} |(\mathbb{M}_n - M)(S) - (\mathbb{M}_n - M)(S_0)| > \inf_{A \in \mathcal{S}_{n,k}} (M(S) - M(S_0)) \right). \quad (5.11)$$

For $c(\gamma) = \min(\gamma - 1/2, 1 - \gamma) > 0$,

$$M(S) - M(S_0) = (\gamma - 1/2)(F(S_0) - F(S_0 \cap S)) + (1 - \gamma)F(S_0^c \cap S) \geq c(\gamma)F(S \triangle S_0),$$

and hence, (5.11) is bounded by

$$\begin{aligned} &\sum_{k=0}^{k_n} P^* \left( \sup_{A \in \mathcal{S}_{n,k}} |(\mathbb{M}_n - M_m)(S) - (\mathbb{M}_n - M_m)(S_0)| > c(\gamma) 2^{k-1} \delta_n \right) \\ &\quad + \sum_{k=0}^{k_n} P^* \left( \sup_{A \in \mathcal{S}_{n,k}} |(M_m - M)(S) - (M_m - M)(S_0)| \geq c(\gamma) 2^{k-1} \delta_n \right). \end{aligned} \quad (5.12)$$

Note that $M_m - M$ is a non-random process and hence, each term in the second sum is either 0 or 1. We now show that the second sum in the above display is eventually zero. For this, we apply Lemma 5.4. Note that

$$\begin{aligned} &\sup_{A \in \mathcal{S}_{n,k}} |(M_m - M)(S) - (M_m - M)(S_0)| \\ &\leq |\Phi_m(0) - 1/2| 2^{k+1} \delta_n + \min(\tilde{c}_0 a_n, 2^{k+1} \delta_n) \\ &\quad + \left| \Phi_m \left( \frac{C_0 \sqrt{m} a_n^p}{\sqrt{1 + \sigma_0^2}} \right) - 1 \right| 2^{k+1} \delta_n + \left| \Phi_m \left( \frac{\sqrt{m} \delta_0}{\sqrt{1 + \sigma_0^2}} \right) - 1 \right| 2^{k+1} \delta_n \\ &\leq 4 \left[ |\Phi_m(0) - 1/2| + \left| \Phi_m \left( \frac{\sqrt{m} \delta_0}{\sqrt{1 + \sigma_0^2}} \right) - 1 \right| \right] 2^{k-1} \delta_n \\ &\quad + \left[ \frac{2 \tilde{c}_0 a_n}{\delta_n} + 4 \left| \Phi_m \left( \frac{C_0 \sqrt{m} a_n^p}{\sqrt{1 + \sigma_0^2}} \right) - 1 \right| \right] 2^{k-1} \delta_n. \end{aligned} \quad (5.13)$$

Hence, it suffices to show that the coefficient of $2^{k-1}\delta_n$ in the above expression is smaller than $c(\gamma)$. To this end, fix $0 < \eta < c(\gamma)/8$. For large $m$,

$$|\Phi_m(0) - 1/2| + \left|\Phi_m\left(\sqrt{m}\delta_0/\sqrt{1+\sigma_0^2}\right) - 1\right| \leq \eta.$$

Choose $c_\eta$ such that $a_n = c_\eta m^{-1/(2p)} > \left[\Phi_m^{-1}(1-\eta)\sqrt{1+\sigma_0^2}/(C_0\sqrt{m})\right]^{1/p}$. For large $n$, the coefficient of $2^{k-1}\delta_n$ in (5.13) is then bounded by

$$8\eta + \frac{\tilde{c}_0 c_\eta}{K_1} < c(\gamma),$$

for $K_1 > (\tilde{c}_0 c_\eta)/(c(\gamma) - 8\eta)$. Hence, each term in the second sum of (5.12) is zero for a suitably large choice of the constant $K_1$. Note that the first term in (5.12) can be written as

$$\sum_{k=0}^{k_n} P^*\left(\sup_{A \in \mathcal{S}_{n,k}} \left|\mathbb{G}_n g_m(\bar{Y}) 1_{A \triangle S_0}(X)\right| > c(\gamma) 2^{k-1}\delta_n\sqrt{n}\right), \tag{5.14}$$

where $g_m(y) = \Phi\left(\sqrt{m}y\right) - \gamma$. We are now in a position to apply Theorem 5.5 to each term of (5.14). In the setup of Theorem 5.5, $N = c(\gamma)2^{k-1}\delta_n\sqrt{n}$. The concerned class of functions is $\mathcal{G}_{n,k} = \{g_m(\bar{Y})1_B(X) : B = A\triangle S_0, B \in \mathcal{S}_{n,k}\}$. Note that $\|g_m 1_B\|_{L_2(P)} \leq [E1_B(X)]^{1/2} \leq (2^{k+1}\delta_n)^{1/2}$. So we can pick $R = R_{n,k} = (2^{k+1}\delta_n)^{1/2}$. As $\mathcal{S}_{n,k} \subset \mathcal{S}$, $N_{[\,]}(u, \{A\triangle S_0 : A \in \mathcal{S}_{n,k}\}, L_2(P)) \leq (N_{[\,]}(u, \mathcal{S}, L_2(P)))^2$ for any $u > 0$. Also, starting with a bracket $[f_L, f_U]$ for $\{A\triangle S_0 : A \in \mathcal{S}_{n,k}\}$ containing $B$ with $\|f_U - f_L\|_{L_2(P)} \leq u$, we can obtain brackets for the class $\mathcal{G}_{n,k}$ using the inequality

$$\Phi\left(\sqrt{m}y\right) f_L - \gamma f_U \leq g_m(y)1_B(x) \leq \Phi\left(\sqrt{m}y\right) f_U - \gamma f_L.$$

As $\|g_m\|_\infty \leq 1$,

$$\|(\Phi\left(\sqrt{m}y\right) f_U - \gamma f_L) - (\Phi\left(\sqrt{m}y\right) f_L - \gamma f_U)\|_{L_2(P)} \leq u.$$

Hence, $H_B(u, \mathcal{G}_{n,k}, L_2(P)) \leq H_B(u, \mathcal{S}, L_2(P))$. Using the fact that in dimension $d$, $H_B(u, \mathcal{S}, L_2(P)) = \log(N_{[\,]}(u, \mathcal{S}, L_2(P))) \leq A_0 u^{-(d-1)}$ for $d \geq 2$ (see Bronšteĭn (1976)), we get

$$H_B(u, \mathcal{G}_{n,k}, L_2(P)) \leq A_0 u^{-1}$$

for some constant $A_0 > 0$ (depending only on the design distribution). The conditions of Theorem 5.5 then translate to

$$\begin{aligned} 2^{k-1} c(\gamma) \delta_n \sqrt{n} &\geq 2 C_2 \max(A_0, 1)(2^{k+1}\delta_n)^{1/4} \\ C_2^2 &\geq C^2(C_3 + 1) \text{ and} \\ c(\gamma) 2^{k-1} \delta_n \sqrt{n} &\leq C_3 \sqrt{n} 2^{k+1} \delta_n. \end{aligned}$$

It can be seen that for $K_1 \geq 2^9 (C_2 \max(A_0, 1)/c(\gamma))^{4/3}$, $C_3 = c(\gamma)/4$ and $C_2 = \sqrt{5}C/2$, these conditions are satisfied, and hence, we can bound (5.14) by

$$\sum_{k=0}^{k_n} C \exp\left\{ \frac{-2^{k-3} c^2(\gamma) \delta_n n}{C^2(C_3 + 1)} \right\}$$

As $\delta_n \gtrsim n^{-2/3}$ (the symbol $\gtrsim$ is used to denote the corresponding $\geq$ inequality holding up to some finite positive constant), the term $\delta_n n$ diverges to $\infty$ as $n \to \infty$. Hence, the above display converges to zero. This completes the proof. $\square$

**Remark 5.7.** *The result also holds for values of $\delta_n$ larger than the one prescribed above. Hence, the above result also gives consistency though it requires $m$ to grow as $m_0 n^{\beta}$. In terms on the total budget, choosing $\beta = 4p/3$ corresponds to the optimal rate. In this case, $\delta_n$ is of the order $n^{-2/3}$ or $N^{-2/(4p+3)}$. This is the minimax rate obtained for a related density level set problem in Tsybakov (1997, Theorem 2) (see also Polonik (1995, Theorem 3.7)).*

Note that the bounds deduced for the two sums in (5.12) depend on $\mu$ only through

$p$ and $\delta_0$, e.g., the exponential bounds from Theorem 5.5 depend on the class of functions only through their entropy and norm of the envelope which do not change with $\mu$. Hence, we have the following result which is similar in flavor to the upper bounds deduced for level-set estimates in Tsybakov (1997).

**Corollary 5.8.** *For the choice of $\delta_n$ given in Proposition 5.6,*

$$\limsup_{n\to\infty} \sup_{\mu\in\mathcal{F}_p} E_\mu^*\left[\delta_n^{-1} d(\hat{S}_n, S_0)\right] < \infty. \tag{5.15}$$

*Here, $E_\mu$ is the expectation with respect to the model with a particular $\mu \in \mathcal{F}_p$. The other features of the model such as error distribution and the design distribution do not change.*

*Proof.* Note that

$$\begin{aligned} E_\mu^*\left[\delta_n^{-1} d(\hat{S}_n, S_0)\right] &\leq 1 + \sum_{k\geq 0, 2^k\delta_n\leq 1} 2^{k+1} P^*\left(2^k < \delta_n^{-1} d(\hat{S}_n, S_0) \leq 2^{k+1}\right) \\ &\leq 1 + \sum_{k\geq 0, 2^k\delta_n\leq 1} 2^{k} P^*\left(\inf_{A\in\mathcal{S}_{n,k}} \mathbb{M}_n(A) - \mathbb{M}_n(S_0) \leq 0\right). \end{aligned}$$

The probabilities $P^*\left(\inf_{A\in\mathcal{S}_{n,k}} \mathbb{M}_n(A) - \mathbb{M}_n(S_0) \leq 0\right)$ can be bounded in an identical manner to that in the proof of the above Proposition and hence, we get

$$\sup_{\mu\in\mathcal{F}_p} E_\mu^*\left[\delta_n^{-1} d(\hat{S}_n, S_0)\right] \leq 1 + \sum_{k=0}^{k_n} C 2^{k+1} \exp\left\{\frac{-2^{k-3} c^2(\gamma)\delta_n n}{C^2(C_3+1)}\right\}.$$

As $\delta_n n \to \infty$, the right side of the above is bounded and hence, we get the result. $\square$

### 5.3.2 Regression Setting

With $\tau_0 = 0$, recall that

$$\mathbb{M}_n(S) = \frac{1}{n} \sum_{k,l: x_{kl}\in\mathcal{I}_n} \left\{\Phi\left(\sqrt{n h_n^2}\hat{\mu}(x_{kl})\right) - \gamma\right\} 1_S(x_{kl}).$$

For any fixed $\gamma \in (1/2, 1)$, it can be shown that $\hat{S}_n$ is consistent for $S_0$, i.e., $d(\hat{S}_n, S_0)$ converges in probability to zero.

**Theorem 5.9.** *Assume $S_0$ to be a closed convex set and the unique minimizer of $M(S)$, where*

$$M(S) = (1/2 - \gamma)\lambda(S_0 \cap S) + (1 - \gamma)\lambda(S_0^c \cap S).$$

*Then, $\sup_{S \in \mathcal{S}} |\mathbb{M}_n(S) - M(S)|$ converges in probability to zero and $\hat{S}_n$ is consistent for $S_0$ in the sense that $d(\hat{S}, S_0)$ converges in probability to zero for any $\gamma \in (0.5, 1)$.*

As was the case in the dose-response setting (see Remark 5.3), a more general result holds and is proved in Section D.2 of Appendix D.

We now deduce a bound on the rate of convergence of $\hat{S}_n$ (for a fixed $\gamma \in (1/2, 1)$). We first consider the population equivalent of $\mathbb{M}_n$, given here by $\bar{M}_n(S) = E\{\mathbb{M}_n(S)\}$ which can be simplified as follows. Let

$$Z_{kl} = \frac{1}{\sqrt{nh_n^2}} \sum_{k',l':x_{kl} \in \mathcal{I}_n}^{n} \epsilon_{k'l'} K\left(\frac{x_{kl} - x_{k'l'}}{h_n}\right),$$

for $k = 1, \ldots, n$, and $Z_0$ be a standard normal random variable independent of $Z_{in}$'s. For notational simplicity, $\sum_{k,l}$ (equivalently, $\sum_{k',l'}$) is used to denote a sum over the set $\{k, l : x_{kl} \in \mathcal{I}_n\}$ unless stated otherwise. Also, let

$$\bar{\mu}(x) = \frac{1}{nh_n^2}\sum_{k',l'} \mu(x_{k'l'})K\left(\frac{x - x_{k'l'}}{h_n}\right) \text{ and } \Sigma_n^2(x) = \frac{1}{nh_n^2}\sum_{k',l'} \sigma_0^2 K^2\left(\frac{x - x_{k'l'}}{h_n}\right). \tag{5.16}$$

Note that $\sqrt{nh_n^2}\hat{\mu}(x_{kl}) = \sqrt{nh_n^2}\bar{\mu}(x_{kl}) + Z_{kl}$ and $\mathrm{Var}(Z_{kl}) = \Sigma_n^2(x_{kl})$. We have

$$\begin{aligned} E\left[\Phi\left(\sqrt{nh_n^2}\hat{\mu}(x_{kl})\right)\right] &= E\left[\Phi\left(\sqrt{nh_n^2}\bar{\mu}(x_{kl}) + Z_{kl}\right)\right] \\ &= E\left[1\left(Z_0 \leq \sqrt{nh_n^2}\bar{\mu}(x_{kl}) + Z_{kl}\right)\right] \\ &= \Phi_{kl,n}\left(\frac{\sqrt{nh_n^2}\bar{\mu}(x_{kl})}{\sqrt{1 + \Sigma_n^2(x_{kl})}}\right), \end{aligned}$$

where $\Phi_{kl,n}$ denotes the distribution function of $(Z_0 - Z_{kl})/\sqrt{1+\Sigma_n^2(x_{kl})}$. For $x_{kl} \in \mathcal{I}_n$, $\Sigma_n^2(x_{kl})$ and $\Phi_{kl,n}$ do not vary with $k$ and $l$ and hence, we denote them by $\tilde{\Sigma}_n^2$ and $\tilde{\Phi}_n$ for convenience. We get

$$\bar{M}_n(S) = E\{\mathbb{M}_n(S)\} = \frac{1}{n}\sum_{k,l}\left\{\tilde{\Phi}_n\left(\frac{\sqrt{nh_n^2}\bar{\mu}(x_{kl})}{\sqrt{1+\tilde{\Sigma}_n^2}}\right) - \gamma\right\} 1_S(x_{kl}). \quad (5.17)$$

Also, for $x_{kl} \in \mathcal{I}_n$, any $\eta > 0$ and sufficiently large $n$,

$$\frac{1}{nh_n^2\tilde{\Sigma}_n^2} \sum_{\substack{k',l' \\ \rho(x_{kl},x_{k'l'})\leq L_0 m h_n}} E\left[\epsilon_{k'l'}^2 K^2\left(\frac{x_{kl}-x_{k'l'}}{h_n}\right) 1\left(\frac{|\epsilon_{k'l'}|K\left((x_{kl}-x_{k'l'})/h_n\right)}{\sqrt{nh_n^2}\tilde{\Sigma}_n} > \eta\right)\right]$$

is bounded by

$$\frac{2\left\lceil 2L_0 m h_n\right\rceil^2 \|K\|_\infty^2}{nh_n^2(\sigma_0^2\bar{K}^2)} E\left[\epsilon_{11}^2 1\left(\frac{2\|K\|_\infty}{nh_n^2(\sigma_0\sqrt{\bar{K}^2})}|\epsilon_{11}| > \eta\right)\right],$$

which converges to zero. Hence, by Lindeberg–Feller central limit theorem, $Z_{kl}/\tilde{\Sigma}_n$ and consequently, $\tilde{\Phi}_n$ converge weakly to $\Phi$. Further, by Pólya's theorem, $\tilde{\Phi}_n$ converges uniformly to $\Phi$ as $n \to \infty$, a fact we use in the proof of Lemma 5.10.

We now consider the distance $d(\hat{S}_n, S_0)$, the rate of convergence of which is driven by the behavior of how small the difference $\mathbb{M}_n - M$ is and how $M$ behaves in the vicinity of $S_0$. We split the difference $\mathbb{M}_n - M$ into $\mathbb{M}_n - \bar{M}_n$ and $\bar{M}_n - M$ and study them separately. We first derive a bound on the distance between $\bar{M}_n$ and $M$.

**Lemma 5.10.** *There exist a positive constant $c_1$ such that for any $a_n \downarrow 0$, $\delta > 0$ and*

$\lambda(S\triangle S_0) < \delta$,

$$\begin{aligned}\left|(\bar{M}_n - M)(S) - (\bar{M}_n - M)(S_0)\right| \le\ & |\tilde{\Phi}_n(0) - 1/2|\delta + \min(c_0 a_n, \delta) \\ & + \left|\tilde{\Phi}_n\left(\frac{\sqrt{nh_n^2}C_0(a_n - 2L_0h_n)^p}{2\sqrt{1+\Sigma^2}}\right) - 1\right|\delta \\ & + \left|\tilde{\Phi}_n\left(\frac{\sqrt{nh_n^2}\delta_0}{2\sqrt{1+\Sigma^2}}\right) - 1\right|\delta + c_1 h_n.\end{aligned} \tag{5.18}$$

*Proof.* Let $Bin_{kl} = \{x = (x_1, x_2) : \ k/m \le x_1 < (k+1)/m, l/m \le x_2 < (l+1)/m\}$. Recall that

$$\begin{aligned}& M(S) - M(S_0) \\ =\ & \int \left\{(1/2)1_{S_0}(x) + 1_{S_0^c}(x) - \gamma\right\}\{1_S(x) - 1_{S_0}(x)\}\,dx \\ =\ & \sum_{0\le k,l\le (m-1)} \int_{Bin_{kl}} \left\{(1/2)1_{S_0}(x) + 1_{S_0^c}(x) - \gamma\right\}\{1_S(x) - 1_{S_0}(x)\}\,dx \\ =\ & \sum_{0\le k,l\le (m-1)} \left[\int_{Bin_{kl}\cap(S_0\cap S^c)} \left\{\frac{1}{2} - \gamma\right\}dx + \int_{Bin_{kl}\cap(S_0^c\cap S)} \{1-\gamma\}\,dx\right] \\ =\ & \sum_{k,l: x_{kl}\in\mathcal{I}_n} \left[\int_{Bin_{kl}\cap(S_0\cap S^c)} \left\{\frac{1}{2} - \gamma\right\}dx + \int_{Bin_{kl}\cap(S_0^c\cap S)} \{1-\gamma\}\,dx\right] + e_n^{(1)}.\end{aligned}$$

Here, $e_n^{(1)}$ is the remainder term arising out of replacing the sum of all choices of $k$ and $l$ to sum over $\{(k,l) : x_{kl} \in \mathcal{I}_n\}$. As the integrands in the above sum are bounded by 1, $|e_n^{(1)}| \le \lambda([0,1]^2 \backslash \mathcal{I}_n) = O(h_n)$. Also,

$$\begin{aligned}\bar{M}_n(S) - \bar{M}_n(S_0) &= \frac{1}{n}\sum_{k,l}\left\{\tilde{\Phi}_n\left(\frac{\sqrt{nh_n^2}\bar{\mu}(x_{kl})}{\sqrt{1+\Sigma_n^2(x_{kl})}}\right) - \gamma\right\}\{1_S(x_{kl}) - 1_{S_0}(x_{kl})\} \\ &= \sum_{k,l}\int_{x\in Bin_{kl}}\left\{\tilde{\Phi}_n\left(\frac{\sqrt{nh_n^2}\bar{\mu}(x_{kl})}{\sqrt{1+\Sigma_n^2(x_{kl})}}\right) - \gamma\right\}\{1_S(x_{kl}) - 1_{S_0}(x_{kl})\}\,dx.\end{aligned}$$

Consequently,

$$
\begin{aligned}
(\bar{M}_n - M)(S) - (\bar{M}_n - M)(S_0) \\
&= \sum_{k,l} \int_{x \in Bin_{kl} \cap (S_0 \cap S^c)} \left( \tilde{\Phi}_n \left( \frac{\sqrt{nh_n^2}\bar{\mu}(x_{kl})}{\sqrt{1+\Sigma_n^2(x_{kl})}} \right) - \frac{1}{2} \right) dx \\
&\quad + \sum_{k,l} \int_{x \in Bin_{kl} \cap (S_0^c \cap S)} \left( \tilde{\Phi}_n \left( \frac{\sqrt{nh_n^2}\bar{\mu}(x_{kl})}{\sqrt{1+\Sigma_n^2(x_{kl})}} \right) - 1 \right) dx \\
&\quad + e_n^{(2)}, \qquad (5.19)
\end{aligned}
$$

where $e_n^{(2)}$, the contribution of the terms at $\partial(S_0 \cap S)$ along with $e_n^{(1)}$, is bounded by

$$
|e_n^{(1)}| + \sum_{k,l: Bin_{kl} \cap \partial(S_0 \cap S) \neq \phi} \int_{Bin_{kl}} 2dx.
$$

This is further bounded by $2\lambda(\{x : \rho(x, \partial S_0) < 2/m\}) + 2\lambda(\{x : \rho(x, \partial S) < 2/m\})$ which is at most $2c_0(2/m) + 2c_0(2/m) = 8c_0/m$ using (5.9) ($\lambda\{x : \rho(x, \partial S) < \alpha\} \leq \lambda(S^\alpha \backslash_\alpha S)$ for any $\alpha > 0$). Hence, for some $\tilde{c}_1 > 0$,

$$
|e_n^{(2)}| \leq O(h_n) + 8c_0/m \leq \tilde{c}_1 h_n.
$$

This contribution is accounted for in the last term of (5.18).

We now study the contribution of the other terms in the right side of (5.19). Note that the integrand in the first sum in the right side of (5.19) is precisely $(\tilde{\Phi}_n(0) - 1/2)$ whenever $Bin_{kl} \subset_{(L_0 h_n)} S_0$ as $\bar{\mu}(x_{kl})$ is zero. As the integrand is also bounded by 1, the first sum in the right side of (5.19) is then bounded by

$$
|\tilde{\Phi}_n(0) - 1/2|\delta + \lambda((S_0 \backslash_{(L_0 h_n)} S_0) \cap S) \leq |\tilde{\Phi}_n(0) - 1/2|\delta + \min(c_0 L_0 h_n, \delta).
$$

Choosing $c_1 = \tilde{c}_1 + c_0 L_0$, the second term on the right side of the above display is also accounted for in the last term in (5.18). Further, let $S_n = \{x : \rho(x, S_0) > a_n\}$.

Note that $\lambda(S_n^c \backslash S_0) \leq c_0 a_n$ using (5.9). Hence, the second sum in (5.19) is bounded by

$$\begin{aligned}
&\int_{S_n^c \backslash S_0} 1dx + \sum_{k,l} \int_{x \in Bin_{kl} \cap (S_n^c \cap S)} \left| \tilde{\Phi}_n \left( \frac{\sqrt{nh_n^2}\bar{\mu}(x_{kl})}{\sqrt{1+\Sigma_n^2(x_{kl})}} \right) - 1 \right| dx \\
\leq \quad & \min(c_0 a_n, \delta) + \sum_{k,l} \int_{x \in Bin_{kl} \cap (S_n \cap S)} \left| \tilde{\Phi}_n \left( \frac{\sqrt{nh_n^2}\bar{\mu}(x_{kl})}{\sqrt{1+\Sigma_n^2(x_{kl})}} \right) - 1 \right| dx,
\end{aligned}$$

To bound the second term in right side of the above, note that as $x_{kl} \in \mathcal{I}_n$,

$$\bar{\mu}(x_{kl}) = E\left[\mu(x_{kl} + h_n Z_n)\right] \left\{ \frac{1}{nh_n^2} \sum_{r,s:|r|,|s| \leq L_0 m h_n} K\left(\left(\frac{r}{m}, \frac{s}{m}\right)\right) \right\},$$

where $Z_n$ is a discrete random variable supported on $\{(r/m, s/m) : |r|, |s| \leq L_0 m h_n\}$ with mass function $P\left[Z_n = (r/m, s/m)\right] \propto K\left((r/m, s/m)\right)$. Hence, the argument of $\tilde{\Phi}_n$ can be written as

$$\sqrt{nh_n^2} E\left[\mu(x_{kl} + h_n Z_n)\right] \frac{\sum_{r,s:|r|,|s| \leq L_0 m h_n} K\left((r/m, s/m)\right)}{nh_n^2 \sqrt{1 + \tilde{\Sigma}_n^2}}$$

Note that

$$\frac{\sum_{r,s:|r|,|s| \leq L_0 m h_n} K\left((r/m, s/m)\right)}{nh_n^2 \sqrt{1 + \tilde{\Sigma}_n^2}} = \frac{1}{\sqrt{1+\Sigma^2}} + o(1),$$

uniformly in $k$ and $l$ for $x_{kl} \in S \cup S_0$. For $x_{kl} \in S_n \cap S_0$ and $a_n < \kappa_0$, when $\rho(x_{kl} + h_n Z_n, S_0) < \kappa_0$, by triangle inequality,

$$\mu(x_{kl} + h_n Z_n) \geq C_0 \rho(x_{kl} + h_n Z_n)^p \geq C_0(\rho(x_{kl}, S_0) - \rho(x_{kl}, x_{kl} + h_n Z_n))^p.$$

As $\rho(x_{kl}, x_{kl} + h_n Z_n) \leq 2L_0 h_n$,

$$\mu(x_{kl} + h_n Z_n) > C_0(a_n - 2L_0 h_n)^p.$$

On the other hand, when $\rho(x_{kl}+h_nZ_n,S_0)\geq\kappa_0$, $\mu(x_{kl}+h_nZ_n)>\delta_0$. Consequently, for $x_{kl}\in S_n\cap S_0$, we get

$$\begin{aligned}\left|\tilde{\Phi}_n\left(\frac{\sqrt{nh_n^2}\bar{\mu}(x_{kl})}{\sqrt{1+\Sigma_n^2(x_{kl})}}\right)-1\right| &\leq \left|\tilde{\Phi}_n\left(\frac{\sqrt{nh_n^2}C_0(a_n-2L_0h_n)^p}{2\sqrt{1+\Sigma^2}}\right)-1\right| \\ &\quad+\left|\tilde{\Phi}_n\left(\frac{\sqrt{nh_n^2}\delta_0}{2\sqrt{1+\Sigma^2}}\right)-1\right|.\end{aligned}$$

As $\lambda(S_n\cap S)<\delta$, we get the result. $\square$

We now consider the term $\mathbb{M}_n(S)-\bar{M}_n(S)$. With $\widetilde{W}_{kl}$s as defined in (5.4), let $W_{kl}=\widetilde{W}_{kl}-E\{\widetilde{W}_{kl}\}$. Then

$$\mathbb{M}_n(S)-\bar{M}_n(S)=\frac{1}{n}\sum_{k,l}W_{kl}1_S(x_{kl}).$$

For notational ease, we define $W_{kl}\equiv 0$ whenever $x_{kl}\notin\mathcal{I}_n$. As the kernel $K$ is compactly supported, $W_{kl}$ is independent of all $W_{k'l'}$s except for those in the set $\{W_{k'l'}:(k',l')\in(1,\ldots,m\}^2,\rho((k,l),(k',l'))\leq 2L_0mh_n\}$. The cardinality of this set is at most $m'=16L_0^2nh_n^2$. Hence, $\{W_{kl}\}_{1\leq k,l\leq m}$ is an $(\sqrt{m'}/2)$-dependent random field. For

$$k_i=i+k\left\lceil\sqrt{m'}\right\rceil,\ l_j=j+l\left\lceil\sqrt{m'}\right\rceil \text{ and } r_{ij}=\sum_{k,l:1\leq k_i,l_j\leq m}1, \tag{5.20}$$

let

$$\begin{aligned}\tilde{d}_n(S_1,S_2)&=\tilde{d}_n(1_{S_1},1_{S_2})\\ &=\left\{\max_{1\leq i,j\leq\lceil\sqrt{m'}\rceil}\left[\frac{1}{r_{ij}}\sum_{k,l:1\leq k_i,l_j\leq m}(1_{S_1}(x_{k_il_j})-1_{S_2}(x_{k_il_j}))^2\right]\right\}^{1/2}\end{aligned}$$

and $\|S\|_n=\|1_S\|_n=\tilde{d}_n(S,\phi)$. Then, the following relation holds.

**Lemma 5.11.** *For sufficiently large $n$,*

$$P\left(\frac{1}{n}\left|\sum_{k,l} W_{kl} 1_S(x_{kl})\right| \geq a\right) \leq 2\exp\left[-\frac{na^2}{16m'\|S\|_n^2}\right].$$

The proof is given in Section D.3 of Appendix D. In fact, such a result holds for general (bounded) $(\sqrt{m'}/2)$-dependent random fields $\{V_{kl} : 1 \leq k, l \leq m\}$ with $|V_{kl}| \leq 1$ and weights $g(x_{kl})$ (instead of $1_S(x_{kl})$'s) as long as $n/m' \to \infty$, i.e., it can be shown that

$$P\left(\frac{1}{n}\left|\sum_{k,l} V_{kl} g(x_{kl})\right| \geq a\right) \leq 2\exp\left[-\frac{na^2}{16m'\|g\|_n^2}\right]. \tag{5.21}$$

Moreover, we can generalize the above to a probability bound on the maxima of an empirical process.

**Theorem 5.12.** *Let $\mathcal{G}$ denote a class of weight functions $g : \{x_{kl} : 1 \leq k, l \leq m\} \mapsto \mathbb{R}$ and $H$ denote the entropy of this class with respect to covering numbers and the metric $\tilde{d}_n$. Assume $\sup_{g\in\mathcal{G}} \|g\|_n \leq R$. Let $V_{kl}$s be random variables with $|V_{kl}| \leq 1$ such that the inequality* (5.21) *holds for all $g \in \mathcal{G}$. Then, there exists a universal constant $C > 0$ such that all $\delta_1 > \delta_2 \geq 0$ satisfying*

$$\sqrt{n/m'}(\delta_1 - \delta_2) \geq C\left(\int_{\delta_2/8}^{R} H^{1/2}(u, \mathcal{G}, \tilde{d}_n)du \vee R\right), \tag{5.22}$$

*we have*

$$P^*\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{k,l} V_{kl} g(x_{kl})\right| \geq \delta\right] \leq C\exp\left[-\frac{n(\delta_1 - \delta_2)^2}{Cm'R^2}\right].$$

The above result states that the supremum of weighted average of (bounded) $(\sqrt{m'}/2)$-dependent random fields, where weights belong to a given class, has sub-gaussian tails. As mentioned earlier, we expect this to be useful in $m$-approximation approaches that are used for deriving limit theorems for dependent random variables and to obtain their empirical process extensions. Here, we use it to control the centered empirical averages $\mathbb{M}_n - \bar{M}_n$. The proof of the above result is outlined in Section D.4 of Appendix D.

We are now in a position to deduce a bound on the rate of convergence of $d(\hat{S}_n, S_0)$.

**Proposition 5.13.** *Let* $\nu_n = \max\left\{h_n, (nh_n^2)^{-1/(2p)}\right\}$*. For some* $K_1 > 0$*, and* $\delta_n = K_1\nu_n$*,* $P^*\left(d(\hat{S}_n, S_0) > \delta_n\right) \to 0$ *as* $n \to \infty$*.*

*Proof.* Let $k_n$ be the smallest integer such that $2^{k_n+1}\delta_n \geq 1$.For $0 \leq k \leq k_n$, let $\mathcal{S}_{n,k} = \left\{S : S \in \mathcal{S},\ 2^k\delta_n < d(S, S_0) \leq 2^{k+1}\delta_n\right\}$. As, $\hat{S}_n$ is the minimizer for $\mathbb{M}_n$,

$$P^*\left(d(\hat{S}_n, S_0) > \delta_n\right) \;\leq\; \sum_{k=0}^{k_n} P^*\left(\inf_{A\in\mathcal{S}_{n,k}} \mathbb{M}_n(A) - \mathbb{M}_n(S_0) \leq 0\right).$$

The sum on the right side can be written as:

$$\sum_{k=0}^{k_n} P^*\left(\sup_{A\in\mathcal{S}_{n,k}} |(\mathbb{M}_n - M)(S) - (\mathbb{M}_n - M)(S_0)| > \inf_{A\in\mathcal{S}_{n,k}} (M(S) - M(S_0))\right). \quad (5.23)$$

For $c(\gamma) = \min(\gamma - 1/2, 1 - \gamma)$, $M(S) - M(S_0) \geq c(\gamma)\lambda(S\triangle S_0)$, and hence (5.24) is bounded by

$$\begin{aligned}
&\sum_{k=0}^{k_n} P^*\left(\sup_{A\in\mathcal{S}_{n,k}} \left|(\mathbb{M}_n - \bar{M}_n)(S) - (\mathbb{M}_n - \bar{M}_n)(S_0)\right| > c(\gamma)2^{k-1}\delta_n\right) \\
&\qquad + \sum_{k=0}^{k_n} 1\left[\sup_{A\in\mathcal{S}_{n,k}} \left|(\bar{M}_n - M)(S) - (\bar{M}_n - M)(S_0)\right| \geq c(\gamma)2^{k-1}\delta_n\right]. \quad (5.24)
\end{aligned}$$

We first apply Lemma 5.10 to the second sum in the above display. Note that

$$\begin{aligned}
&\sup_{A\in\mathcal{S}_{n,k}} \left|(\bar{M}_n - M)(S) - (\bar{M}_n - M)(S_0)\right| \\
&\leq\ |\tilde{\Phi}_n(0) - 1/2|2^{k+1}\delta_n + \min(h_n, 2^{k+1}\delta_n) \\
&\quad + \left|\tilde{\Phi}_n\left(\frac{\sqrt{nh_n^2}C_0(a_n - 2L_0h_n)^p}{2\sqrt{1+\Sigma^2}}\right) - 1\right| 2^{k+1}\delta_n
\end{aligned}$$

$$
\begin{aligned}
& + \left| \tilde{\Phi}_n \left( \frac{\sqrt{n h_n^2}\delta_0}{2\sqrt{1+\Sigma^2}} \right) - 1 \right| 2^{k+1}\delta_n + c_1 h_n \\
\leq \; & 4 \left[ |\tilde{\Phi}_n(0) - 1/2| + \left| \tilde{\Phi}_n \left( \frac{\sqrt{n h_n^2}\delta_0}{2\sqrt{1+\Sigma^2}} \right) - 1 \right| \right] 2^{k-1}\delta_n \\
& + \left[ \frac{2c_0 a_n + c_1 h_n}{\delta_n} + 4 \left| \tilde{\Phi}_n \left( \frac{\sqrt{n h_n^2} C_0 (a_n - 2L_0 h_n)^p}{2\sqrt{1+\Sigma^2}} \right) - 1 \right| \right] 2^{k-1}\delta_n.
\end{aligned}
$$

Fix $0 < \eta < c(\gamma)/8$. For large $n$, $|\tilde{\Phi}_n(0) - 1/2| + c_1/(\sqrt{n}\delta_n) < \eta$. Choose $c_\eta$ such that $a_n = c_\eta \nu_n > \left[ 2\tilde{\Phi}_n^{-1}(1-\eta)\sqrt{1+\Sigma^2}/(C_0\sqrt{n h_n^2}) \right]^{1/p} + (2L_0)h_n$. Then the coefficient of $2^{k-1}\delta_n$ in the above display is bounded by

$$
8\eta + \frac{2c_0 c_\eta + c_1}{K_1} < c(\gamma),
$$

when $K_1 > (2c_0 c_\eta + c_1)/(c(\gamma) - 8\eta)$. Hence, for a suitably large choice of $K_1$ each term in the second sum of (5.24) is zero.

We now apply Theorem 5.12 to each term in the first sum of (5.24). For this we use the following claim to obtain a bound on the entropy of the class $\mathcal{S}_{n,k}$.

**Claim A.** We claim that $\sup_{S_1,S_2 \in \mathcal{S}} |\tilde{d}_n^2(S_1, S_2) - \lambda(S_1 \triangle S_2)| = O(h_n)$ and that $H(u, \{B \triangle S_0 : B \in \mathcal{S}_{n,k}\}, \tilde{d}_n) \leq A_1(u - c_2 h_n)^{-1}$ for constants $c_2 > 0$ and $A_1 > 0$. We first use the above claim to prove the result. As a consequence of Claim A, $\sup_{A \in \{B \triangle S_0 : B \in \mathcal{S}_{n,k}\}} \|A\|_n \leq R_{n,k} := (2^{k+1}\delta_n + c_3 h_n)^{1/2}$, for some $c_3 > 0$. Using Theorem 5.12 with $\delta_1 = c(\gamma)2^{k-1}\delta_n$, $\delta_2 = 8c_2 h_n$, we arrive at the condition

$$
\sqrt{n/m'}(c(\gamma)2^{k-1}\delta_n - 8c_2 h_n) \gtrsim (R_{n,k} + c_4 h_n)^{1/2} \vee R_{n,k},
$$

for some $c_4 > 0$. As $\delta_n \gtrsim \nu_n$, this translates to $(2^{k-1}c(\gamma))^4 \delta_n^3 \gtrsim (2^k + c_5)h_n^4$ for some $c_5 > 0$. This holds for all $k$ when $\delta \gtrsim h_n^{4/3}$ which is true as $\delta_n \gtrsim h_n$. Hence, we can bound the first sum in (5.24) by

$$\sum_{k=0}^{k_n} 5C \exp\left[-\frac{n\left(c(\gamma)2^{k-1}\delta_n - 8c_2h_n\right)^2}{Cm'(2^{k+1}\delta_n + c_3h_n)^{1/2}}\right]. \tag{5.25}$$

Consequently, the display in (5.25) is bounded by

$$\sum_{k=0}^{\infty} 5C \exp\left[-\frac{\left(c(\gamma)2^{k-1} - c_6\right)^2 \delta_n}{Ch_n^2(2^{k+1} + c_7)}\right],$$

for some constants $c_6$, $c_7 > 0$. As $\delta_n/h_n^2 \gtrsim h_n^{-1} \to \infty$, we get the result.

*Proof of* **Claim A**. Note that $\tilde{d}^2(S_1, S_2) = \tilde{d}^2(S_1 \triangle S_2, \phi) = \max_{1 \le i,j \le \lceil \sqrt{m'} \rceil} Q_n^{ij}(S_1 \triangle S_2)$, where $Q_n^{ij}$ is the discrete uniform measure on the points $\{x_{k_i l_j} : k_i = i + k\lceil \sqrt{m'} \rceil \le m,\ l_j = j + l\lceil \sqrt{m'} \rceil \le m\}$. Note that each $Q_n^{ij}$ approximates Lebesgue measure at resolution of rectangles of length $m/m' = O(h_n)$. The rectangles that intersect with the boundary of a set $S$ account for the difference $|Q_n^{ij}(S) - \lambda(S)|$. As argued in the proof of Lemma 5.10, the error $\sup_{S \in \mathcal{S}} \max_{i,j} |Q^{ij}(S) - \lambda(S)| \le \lambda(\{x : \rho(x, \partial S) < O(h_n)\}$, which is $O(h_n)$ using (5.9).

To see that $H(u, \{B \triangle S_0 : B \in \mathcal{S}_{n,k}\}, \tilde{d}_n) \le A_1(u - c_2h_n)^{-1}$, first, note that $H(u, \{B \triangle S_0 : B \in \mathcal{S}_{n,k}\}, \tilde{d}_n) \le H(u, \mathcal{S}, \tilde{d}_n)$. For any convex set $S$, it can be shown from arguments analogous to those in the proof for Lemma 5.10 that for some $c_2 > 0$,

$$\max_{1 \le i,j \le \lceil \sqrt{m'} \rceil} Q^{ij}(S^{\delta} \backslash_{\delta} S) \le \lambda(S^{(\delta + c_2h_n)} \backslash_{(\delta + c_2h_n)} S) \le c_0(\delta + c_2h_n).$$

If $S_1, \ldots, S_r$ are the center of the Hausdorff balls with radius $\delta$ that cover $\mathcal{S}$ (see (D.1) in Appendix D for a definition of Hausdorff distance $d_H$), then $[_{\delta}S_i, S_i^{\delta}]$, $i \le r$ form brackets that cover $\mathcal{S}$. The sizes of these brackets are $(c_0(\delta + c_2h_n))^{1/2}$ in terms of the distance $\tilde{d}_n$. Hence,

$$H((c_0(\delta + c_2 h_n))^{1/2}, \mathcal{S}, \tilde{d}_n) \leq H_B((c_0(\delta + c_2 h_n))^{1/2}, \mathcal{S}, \tilde{d}_n) \leq H(\delta, \mathcal{S}, d_H).$$

Letting $u = c_0(\delta + c_2 h_n))^{1/2}$ and using the fact that $H(\delta, \mathcal{S}, d_H) \lesssim \delta^{-1/2}$ we get **Claim A**. $\square$

As was the case with Corollary 5.8, Proposition 5.13 extends to the following result in an identical manner.

**Corollary 5.14.** *For the choice of $\delta_n$ given in Proposition 5.13,*

$$\limsup_{n\to\infty} \sup_{\mu\in\mathcal{F}_p} E^*_\mu \left[\delta_n^{-1} d(\hat{S}_n, S_0)\right] < \infty. \tag{5.26}$$

**Remark 5.15.** *The best rate at which the distance $d(\hat{S}_n, S_0)$ goes to zero corresponds to $h_n \sim (nh_n)^{-1/(2p)}$ which yields $\nu_n \sim h_n = h_0 n^{-1/2(p+1)}$. This is slower than the rate we deduced in the dose-response setting in terms of the total budget ($N^{-2/(4p+3)}$). The difference in the rate from the dose-response setting is accounted for by the bias in the smoothed kernel estimates. The regression setting is approximately equivalent to a dose-response model having $(2L_0 h_n)^{-2}$ (effectively) independent covariate observations and $n(2L_0 h_n)^2$ (biased) replications. These replications correspond to the number of observations used to compute $\hat{\mu}$ at a point. If we compare Lemmas 5.4 and 5.10, these biased replications add an additional term of order $h_n$ which is absent in the dose-response setting. This puts a lower bound on the rate at which the set $S_0$ can be approximated. In contrast, the rates coincide for the dose-response and the regression settings in the one-dimensional case; see Chapters 3 and 4. This is due to the fact that in one dimension, the bias to standard deviation ratio ($h_n/(1/\sqrt{nh})$) is of smaller order compared to that in two dimensions ($h_n/(1/\sqrt{nh^2})$) for estimating $\hat{\mu}$. In a nutshell, the curse of dimensionality kicks in at dimension 2 itself in this problem.*

### 5.3.3 Extension to the case of an unknown $\tau_0$

While we deduced our results under the assumption of a known $\tau_0$, in real applications $\tau_0$ is generally unknown. As was the case with the one-dimensional problem, quite a few extensions are possible in this situation. For example, in dose-response setting, if $S_0$ can be safely assumed to contain a positive $F$-measure set $U$, then a simple averaging of the $\bar{Y}$ values realized for $X$'s in $U$ would yield a $\sqrt{mn}$-consistent estimator of $\tau_0$. If a proper choice of $U$ is not available, one can obtain an initial estimate of $\tau_0$ in the dose–response setting as

$$\hat{\tau}_{init} = \operatorname*{argmin}_{\tau \in \mathbb{R}} \mathbb{P}_n \left[ \Phi\left(\sqrt{m}(\bar{Y} - \tau)\right) - \frac{1}{2} \right]^2 .$$

This provides a consistent estimate of $\tau_0$ under mild assumptions. A $\sqrt{mn}$-consistent estimate of $\tau_0$ can then be found by using $\hat{\tau}_{init}$ to compute $\hat{S}_n$ and then averaging the $\bar{Y}$ value for the $X$'s realized in ${}_\delta\hat{S}_n$ for a small $\delta > 0$. Note that this leads to an iterative procedure where this new estimate of $\tau$ is used to update the estimate of $\hat{S}_n$. It can be shown that the rate of convergence remains unchanged if one imputes a $\sqrt{mn}$-consistent estimate of $\tau_0$. A brief sketch of the following result is given in Section D.5.

**Proposition 5.16.** *Let $\hat{S}_n$ now denote the minimizer of*

$$\mathbb{M}_n(S, \hat{\tau}) = \mathbb{P}_n \left[ \left\{ \Phi\left(\sqrt{m}(\bar{Y} - \hat{\tau})\right) - \gamma \right\} 1_S(X) \right] ,$$

*where $\sqrt{mn}\,(\hat{\tau} - \tau_0) = O_p(1)$. For $m = m_0 n^{\beta}$ and $\delta_n$ as defined in Proposition 5.6, we have $P\left(d(\hat{S}_n, S_n) > \delta_n\right) \to 0$.*

In the regression setting as well, an initial consistent estimate of $\tau_0$ can be computed as

$$\hat{\tau}_{init} = \operatorname*{argmin}_{\tau \in \mathbb{R}} \frac{1}{n} \sum_{k,l} \left[ \Phi\left( \sqrt{n h_n^2}(\hat{\mu}(x_{kl}) - \tau) \right) - \frac{1}{2} \right]^2$$

which can then be used to yield a $\sqrt{n}$-consistent estimate of $\tau_0$ using the iterative approach mentioned above. We have the following result for the rate of convergence of $\hat{S}_n$ in the regression setting.

**Proposition 5.17.** *Let $\hat{S}_n$ now denote the minimizer of*

$$\mathbb{M}_n(S, \hat{\tau}) = \frac{1}{n} \sum_{k,l} \left[ \left\{ \Phi \left( \sqrt{n h_n^2} (\hat{\mu}(x_{kl}) - \hat{\tau}) \right) - \gamma \right\} 1_S(X) \right] ,$$

*where $\sqrt{n}\,(\hat{\tau} - \tau_0) = O_p(1)$. For $\delta_n$ as defined in Proposition 5.13, $P\left( d(\hat{S}_n, S_n) > \delta_n \right) \to 0$.*

The proof is outlined in Section D.6 of Appendix D.

## 5.4 Discussion

*Extensions to non-convex baseline sets.* Although we essentially address the situation where the baseline set is convex for dimension $d = 2$, our approach extends past convexity and the two-dimensional setting in presence of an efficient algorithm and for suitable collection of sets. For example, let $\tilde{\mathcal{S}}$ denote such a collection of subsets of $[0, 1]^d$ sets such that

$$\tilde{S}_n = \operatorname*{argmin}_{S \in \tilde{\mathcal{S}}} \mathbb{M}_n(S)$$

is easy to compute. Here, $\mu$ is a real-valued function from $[0, 1]^d$ and $S_0 = \mu^{-1}(\tau_0)$ is assumed to belong to the class $\tilde{\mathcal{S}}$. Then the estimator $\tilde{S}_n$ has the following properties in the dose-response setting.

**Proposition 5.18.** *Assume that $S_0$ is the unique minimizer (up to $F$-null sets) of the population criterion function $M_F$ defined in* (5.8)*. Then $d_F(\tilde{S}_n, S_0)$ converges in probability to zero. Moreover, assume that there exists a constant $\bar{c} > 0$ such that*

*$F(S^\epsilon \backslash_\epsilon S) \leq \bar{c}\epsilon$ for any $\epsilon > 0$ and $S \in \tilde{\mathcal{S}}$, and*

$$H_B(u, \tilde{\mathcal{S}}, L_2(P)) \lesssim u^{-r} \text{ for some } r < 2.$$

*Then, $P\left(d_F(\tilde{S}_n, S_0) > \tilde{\delta}_n\right)$ converges to zero where $\tilde{\delta}_n = K_1 \max(n^{-2/(2+r)}, m^{-1/(2p)})$ for some $K_1 > 0$.*

The proof follows along lines identical to that for Proposition 5.6. Note that the relation of the type $F(S^\epsilon \backslash_\epsilon S) \leq \bar{c}\epsilon$ was needed to derive Lemma 5.4. This assumption simply rules out the sets with highly irregular or non-rectifiable boundaries. Also, the dependence of the rate on the dimension typically comes through $r$ which usually grows with $d$. A similar result can be established in the regression setting as well.

*Connection with level-set approaches.* Note that minimizing $\mathbb{M}_n(S)$ in the dose-response setting is equivalent to minimizing

$$\begin{aligned}
\tilde{\mathbb{M}}_n(S) &= \mathbb{M}_n(S) - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{1}{4} - p_{m,n}(X_i)\right) \\
&= \sum_{i=1}^{n} \frac{1/4 - p_{m,n}(X_i)}{2}\left[1(X_i \in S) - 1(X_i \in S^c)\right].
\end{aligned}$$

This form is very similar to an empirical risk criterion function that is used in Willett and Nowak (2007, equation (7)) in the context of a level-set estimation procedure. It can be deduced that our baseline detection approach ends up finding the level set $S_m = \{x : E\left[p_{m,n}(x)\right] > 1/4\}$ from i.i.d. data $\{p_{m,n}(X_i), X_i\}_{i=1}^n$ with $0 \leq p_{m,n}(X_i) \leq 1$. As $m \to \infty$, $S_m$'s decrease to $S_0$, which is the target set. Hence, any level-set approach could be applied to transformed data $\{p_{m,n}(X_i), X_i\}_{i=1}^n$ to yield an estimate for $S_m$ which would be consistent for $S_0$. Moreover, a similar connection between the two approaches can be made for the regression setting, however the i.i.d. flavor of the observations present in the dose-response setting is lost as $\{p_n(x_{kl})\}_{1\leq k,l \leq m}$ are dependent. While the algorithm from Willett and Nowak (2007) can be implemented to construct the baseline set estimate, it is far from clear how the theoretical properties would then translate to our setting given the dependence of the target function $E\left[p_{m,n}(x)\right]$ on $m$ in the dose-response setting and the dependent nature of the transformed data in the regression setting.

In Scott and Davenport (2007), the approach to the level set estimation problem, using the criterion in Willett and Nowak (2007), is shown to be equivalent to a *cost-sensitive classification* problem. This problem involves random variables $(X, Y, C) \in \mathbb{R}^d \times \{0,1\} \times \mathbb{R}$, where $X$ is a feature, $Y$ a class and $C$ is the cost for misclassifying $X$ when the true label is $Y$. Cost sensitive classification seeks to minimize the expected cost

$$R(G) = E(C\ 1(G(X) \neq Y)), \tag{5.27}$$

where $G$, with a little abuse of notation, refers both to a subset of $\mathbb{R}^d$ and $G(x) = 1(x \in G)$. With $C = |\gamma - Y|$ and $\tilde{Y} = 1(Y \geq \gamma)$, the objective of the cost-sensitive classification, based on $(X, \tilde{Y}, C)$, can be shown to be equivalent to minimizing the excess risk criterion in Willett and Nowak (2007). So, approaches like support vector machines (SVM) and $k$-nearest neighbors ($k$-NN), which can be tailored to solve the cost-sensitive classification problem (see Scott and Davenport (2007)), are relevant to estimating level sets, and thus provide alternative ways to solve the baseline set detection problem. Since the *loss function* in (5.27) is not smooth, one might prefer to work with its surrogates. Some results in this direction can be found in Scott (2011).

*Adaptivity.* We have assumed knowledge of the order of the regularity $p$ of $\mu$ at $\partial S_0$, which is required to achieve the optimal rate of convergence, though not for consistency. The knowledge of $p$ dictates the allocation between $m$ and $n$ in the dose-response setting and the choice of the bandwidth $h_n$ in the regression setting for attaining the best possible rates. When $p$ is unknown, the adaptive properties of dyadic trees (see Willett and Nowak (2007) and Singh et al. (2009)) could conceivably be utilized to develop a near-optimal approach. However, this is a hard open problem and will be a topic of future research.

# Part II

# Multi-stage procedures

# CHAPTER 6

# A generic approach to multistage procedures

Multi-stage procedures, obtained by splitting the sampling budget suitably across stages, and designing the sampling at a particular stage based on information about the parameter obtained from previous stages, have received some attention in recent times (Lan et al., 2009; Tang et al., 2011; Belitser et al., 2013). They are found advantageous over their one-stage counterparts from the perspective of inference. For example, Lan et al. (2009) considered the problem of estimating the *change point* $d_0$ from a regression model $Y = f(X) + \epsilon$ with $f(x) = \alpha_0 1(x \leq d_0) + \beta_0 1(x > d_0)$, $\alpha_0 \neq \beta_0$ and showed that a two-stage estimate converges to $d_0$ at a rate much faster (almost $n$ times) than the estimate obtained from a one-stage approach. In a non-parametric isotonic regression framework, $Y = r(X) + \epsilon$ with $r$ monotone, Tang et al. (2011) achieve a $\sqrt{n}$-rate of convergence (seen usually in parametric settings) for estimating thresholds $d_0$ of type $d_0 = r^{-1}(t_0)$ (for a fixed known $t_0$) by doing a linear approximation at the second stage of sampling. This is a marked improvement over the usual one-stage estimate which converges at the rate $n^{1/3}$. Further, Belitser et al. (2013) considered the problem of estimating the location and size of the maximum of a multivariate regression function, where they avoided the curse of dimensionality through a two-stage procedure.

In the problems mentioned above, a common (multistage) sampling scheme was implemented which we state below.

1. In the first stage, utilize a fixed portion of the design budget to obtain an initial estimate, say, of the location $d_0$ and the nuisance parameters present in the model.

2. Sample the second stage design points in a shrinking neighborhood around the first stage estimator and use the earlier estimation approach (or a different one that leverages on the local behavior of the regression function in the vicinity of $d_0$) to obtain the final estimate of $d_0$ in this "zoomed-in" neighborhood.

This type of an approach adds an extra level of complication as the second stage data is no longer i.i.d. This is due to the dependence of the design points on the first stage estimate of $d_0$. Moreover, in several cases, the second stage estimates are usually constructed by minimizing (or maximizing) a related empirical process sometimes over a random set based on the first stage estimates. In the problems mentioned above, these intricacies were addressed using fairly different theoretical tools starting from first principles. However, in a variety of problems similar in flavor to those mentioned above, a unified approach is possible which we develop in this chapter.

In this chapter, we extend empirical process results originally developed for the i.i.d. setting to situations with dependence of the nature discussed above. In particular, we establish general results for deriving rate of convergence, proving tightness of empirical process and deducing limiting distribution in general multi-stage problems (see Section 6.1). We implement our results on problems from change-point analysis (Section 6.2), inverse isotonic regression (Section 6.3), classification (Section 6.3.1) and mode estimation (Section 6.4).

Our results are also relevant to situations where certain extra/nuisance parameters are estimated from separate data and argmax/argmin functionals of the empirical process acting on functions involving these estimated parameters are considered. We note here that van der Vaart and Wellner (2007) considered similar problems where they provided sufficient conditions for replacing such estimated parameters by their true values, in the sense that $\sup_{d\in\mathcal{D}}\left|\mathbb{G}_n(f_{d,\hat{\theta}} - f_{d,\theta_0})\right|$ converges in probability to zero. Here, $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$, with $\mathbb{P}_n$ denoting the empirical measure, $f_{d,\theta}$ are measurable functions indexed by $(d,\theta) \in \mathcal{D}\times\Theta$ and $\hat{\theta}$ denotes a suitable estimate of the nuisance parameter $\theta_0$. We show that a result of the above form does not generally hold for our examples, (see Proposition 6.7), but the final limit distribution still has a form with estimated nuisance parameters replaced by their true values.

## 6.1 Formulation and general results

A typical two-stage procedure involves estimating certain parameters, say a vector $\theta_n$, from the first stage sample. Let $\hat{\theta}_n$ denote this first stage estimate. Based on $\hat{\theta}_n$, a suitable sampling design is chosen to obtain the second stage estimate of the parameter of interest $d_0$ by minimizing (or maximizing) a criterion function $\mathbb{M}_n(d,\hat{\theta}_n)$ over domain $\mathcal{D}_{\hat{\theta}_n} \subset \mathcal{D}$, i.e.,

$$\hat{d}_n = \operatorname*{argmin}_{d\in\mathcal{D}_{\hat{\theta}_n}} \mathbb{M}_n(d,\hat{\theta}_n). \tag{6.1}$$

We denote the domain of optimization for a generic $\theta$ by $\mathcal{D}_\theta$. We will impose more structure on $\mathbb{M}_n$ as and when needed. We start with a general theorem about deducing the rate of convergence of $\hat{d}_n$ arising from such criterion. In what follows, $M_n$ is typically a population equivalent of the criterion function $\mathbb{M}_n$, e.g., $M_n(d,\theta_n) = E\left[\mathbb{M}_n(d,\theta_n)\right]$, which is at its minimum at the parameter of interest $d_0$ or at a quantity $d_n$ asymptotically close to $d_0$.

**Theorem 6.1.** *Let $\{\mathbb{M}_n(d,\theta),\ n\geq 1\}$ be stochastic processes and $\{M_n(d,\theta),\ n\geq 1\}$ be deterministic functions, indexed by $d\in\mathcal{D}$ and $\theta\in\Theta$. Let $d_n\in\mathcal{D}$, $\theta_n\in\Theta$ and $d\mapsto\rho_n(d,d_n)$ be a measurable map from $\mathcal{D}$ to $[0,\infty)$. Let $\hat{d}_n$ be a (measurable) point* *of minimum of* $\mathbb{M}_n(d, \hat{\theta}_n)$ *over* $d \in \mathcal{D}_{\hat{\theta}_n} \subset \mathcal{D}$*, where* $\hat{\theta}_n$ *is a random map independent of the process* $\mathbb{M}_n(d, \theta)$*. For each* $\tau > 0$ *and some* $\kappa_n > 0$ *(not depending on* $\tau$*), suppose that the following hold:*

*(a) There exists a sequence of sets* $\Theta_n^\tau$ *in* $\Theta$ *such that* $P[\hat{\theta}_n \notin \Theta_n^\tau] < \tau$*.*

*(b) There exist constants* $c_\tau > 0$*,* $N_\tau \in \mathbb{N}$ *such that for all* $\theta \in \Theta_n^\tau$*,* $d \in \mathcal{D}_\theta$ *with* $\rho_n(d, d_n) < \kappa_n$*, and* $n > N_\tau$*,*

$$M_n(d, \theta) - M_n(d_n, \theta) \geq c_\tau \rho_n^2(d, d_n). \tag{6.2}$$

*Also, for any* $\delta \in (0, \kappa_n)$ *and* $n > N_\tau$*,*

$$\sup_{\theta \in \Theta_n^\tau} E^* \sup_{\substack{\rho_n(d, d_n) < \delta, \\ d \in \mathcal{D}_\theta}} |(\mathbb{M}_n(d, \theta) - M_n(d, \theta)) - (\mathbb{M}_n(d_n, \theta) - M_n(d_n, \theta))| \leq C_\tau \frac{\phi_n(\delta)}{\sqrt{n}}, \tag{6.3}$$

*for a constant* $C_\tau > 0$ *and functions* $\phi_n$ *(not depending on* $\tau$*) such that* $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ *is decreasing for some* $\alpha < 2$*.*

*Suppose that* $r_n$ *satisfies*

$$r_n^2 \, \phi_n\left(\frac{1}{r_n}\right) \lesssim \sqrt{n},$$

*and* $P\left(\rho_n(\hat{d}_n, d_n) \geq \kappa_n\right)$ *converges in probability to zero, then* $r_n \, \rho_n(\hat{d}_n, d_n) = O_p(1)$*.*

*Further, if the assumptions in part (b) of the above theorem hold for all sequences* $\kappa_n > 0$ *in the sense that there exist constants* $c_\tau > 0$*,* $C_\tau > 0$*,* $N_\tau \in \mathbb{N}$ *such that for all* $\theta \in \Theta_n^\tau$*,* $d \in \mathcal{D}_\theta$*,* $\delta > 0$ *and* $n > N_\tau$*, (6.2) and (6.3) hold, then justifying the convergence of* $P\left(\rho_n(\hat{d}_n, d_n) \geq \kappa_n\right)$ *to zero is not necessary.*

A version of this result involving a fixed $\kappa_n \equiv \kappa > 0$ also holds where $N_\tau$ is allowed to depend on $\kappa$. The proof uses shelling arguments similar to those in the proof of

Theorem 3.2.5 in van der Vaart and Wellner (1996). It is given in Section E.1 of Appendix E. An intermediate step to applying the above result involves justifying the convergence of $P\left(\rho_n(\hat{d}_n, d_n) \geq \kappa_n\right)$ to zero. As mentioned in the result, if the assumptions in part (b) of the above theorem hold for all sequences $\kappa_n > 0$, then justifying this condition is *not* necessary. This is the case with *most* of the examples that we study in this chapter. The following result is used otherwise.

**Lemma 6.2.** *Let $\mathbb{M}_n$, $M_n$ and $\rho_n$ be as defined in Theorem 6.1. For any fixed $\tau > 0$, let*

$$c_n^\tau(\kappa_n) = \inf_{\rho_n(d,d_n)\geq\kappa_n, d\in\mathcal{D}_\theta} \{M_n(d,\theta) - M_n(d_n,\theta)\}$$

*Suppose that*

$$\sup_{\theta\in\Theta_n^\tau} P\left(2\sup_{d\in\mathcal{D}_\theta} |\mathbb{M}_n(d,\theta) - M_n(d,\theta)| \geq c_n^\tau(\kappa_n)\right) \to 0. \tag{6.4}$$

*Then, $P\left(\rho_n(\hat{d}_n, d_n) \geq \kappa_n\right)$ converges to zero .*

Condition (6.4) requires $c_n^\tau(\kappa_n)$ to be positive (eventually) which ensures that $d_n$ is the unique minimizer of $M_n(d,\theta)$ over the set $d \in \mathcal{D}_\theta$. The proof is given in Section E.2 of Appendix E.

The conclusion of Theorem 6.1, $r_n\ \rho_n(\hat{d}_n, d_n) = O_p(1)$, typically leads to a result of the form $s_n(\hat{d}_n - d_n) = O_p(1)$, $s_n \to \infty$. Once such a result has been established, the next step is to study the limiting behavior of the local process

$$Z_n(h,\hat{\theta}_n) = v_n\left[\mathbb{M}_n\left(d_n + \frac{h}{s_n}, \hat{\theta}_n\right) - \mathbb{M}_n\left(d_n, \hat{\theta}_n\right)\right]$$

for a properly chosen $v_n$. Note that

$$s_n(\hat{d}_n - d_n) = \operatorname*{argmin}_{h: d_n + h/s_n \in \mathcal{D}_{\hat{\theta}_n}} Z_n(h,\hat{\theta}_n).$$

Note that $Z_n$ can be defined in such a manner so that the right hand side is the minimizer of $Z_n$ over the entire domain. To see this, let $\mathcal{D}_{\hat{\theta}_n} = [a_n(\hat{\theta}_n), b_n(\hat{\theta}_n)]$, say (in one dimension). If we extend the definition of $Z_n$ to the entire line by defining

$$Z_n(h, \hat{\theta}_n) = \begin{cases} Z_n(s_n(b_n(\hat{\theta}_n) - d_n)) & \text{for } h > s_n(b_n(\hat{\theta}_n) - d_n) \text{ and} \\ Z_n(s_n(a_n(\hat{\theta}_n) - d_n)) & \text{for } h < s_n(a_n(\hat{\theta}_n) - d_n), \end{cases} \tag{6.5}$$

then, clearly:

$$s_n(\hat{d}_n - d_n) = \operatorname*{argmin}_{\mathbb{R}} Z_n(h, \hat{\theta}_n)\,.$$

In $p$ dimensions, define $Z_n$ outside of the real domain, the translated $\hat{D}_{\hat{\theta}_n}$, to be the supremum of the process $Z_n$ on its real domain. Then the infimum of $Z_n$ over the entire space is also the infimum over the real domain. Such an extension then allows us to apply the argmin continuous mapping theorem (Kim and Pollard, 1990, Theorem 2.7) to arrive at the limiting distribution of $s_n(\hat{d}_n - d_n)$.

In our examples and numerous others, $Z_n$ can be expressed as an empirical process acting on a class of functions changing with $n$, indexed by the parameter $h$ over which the argmax/argmin functional is applied and by the parameter $\theta$ which gets estimated from the first stage data, e.g.,

$$Z_n(h, \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f_{n,h,\theta}(V_i) = \mathbb{G}_n f_{n,h,\theta} + \zeta_n(h, \theta). \tag{6.6}$$

Here, $V_i \sim P$ are i.i.d. random vectors, $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ and $\zeta_n(h, \theta) = \sqrt{n} P f_{n,h,\theta}$ with $\mathbb{P}_n$ denoting the empirical measure induced by $V_i$s. The parameter $\theta$ could be multi-dimensional and would account for the nuisance/design parameters which are estimated from the first stage sample. Moreover, $f_n$'s need not have zero mean and $\sqrt{n} P f_{n,h,\theta}$ could possibly contribute to the drift of the limiting process. First, we provide sufficient conditions for the asymptotic tightness of the processes $Z_n(h, \hat{\theta}_n)$.

**Theorem 6.3.** *Let $\hat{\theta}_n$ be a random variable taking values in $\Theta$ which is independent of the process $Z_n$ defined in (6.6). As in Theorem 6.1, let there exist a (non-random) set $\Theta_n^\tau \subset \Theta$ such that $P[\hat{\theta}_n \notin \Theta_n^\tau] < \tau$, for any fixed $\tau > 0$. For each $\theta \in \Theta$, let $\mathcal{F}_{n,\theta} = \{f_{n,h,\theta} : h \in \mathcal{H}\}$ with measurable envelopes $F_{n,\theta}$. Let $\mathcal{H}$ be totally bounded with respect to a semimetric $\tilde{\rho}$. Assume that for each $\tau, \eta > 0$ and every $\delta_n \to 0$,*

$$\begin{aligned}
\sup_{\theta \in \Theta_n^\tau} PF_{n,\theta}^2 &= O(1), && (6.7)\\
\sup_{\theta \in \Theta_n^\tau} PF_{n,\theta}^2 1\left[F_{n,\theta} > \eta\sqrt{n}\right] &\to 0 && (6.8)\\
\sup_{\substack{\theta \in \Theta_n^\tau \\ \tilde{\rho}(h_1,h_2)<\delta_n}} P(f_{n,h_1,\theta} - f_{n,h_2,\theta})^2 &\to 0 \text{ and} && (6.9)\\
\sup_{\substack{\theta \in \Theta_n^\tau \\ \tilde{\rho}(h_1,h_2)<\delta_n}} |\zeta_n(h_1,\theta) - \zeta_n(h_2,\theta)| &\to 0. && (6.10)
\end{aligned}$$

*Assume, for $\delta > 0$, $\mathcal{F}_{n,\delta} = \{f_{n,h_1,\hat{\theta}} - f_{n,h_2,\hat{\theta}} : \tilde{\rho}(h_1,h_2) < \delta\}$ is suitably measurable (explained below), for each $\theta \in \Theta_n^\tau$, $\mathcal{F}_{n,\theta,\delta}^2 = \{(f_{n,h_1,\theta} - f_{n,h_2,\theta})^2 : \tilde{\rho}(h_1,h_2) < \delta\}$ is $P$-measurable, and*

$$\sup_{\theta \in \Theta_n^\tau} \int_0^\infty \sup_Q \sqrt{\log N\left(u\|F_{n,\theta}\|_{L_2(Q)}, \mathcal{F}_{n,\theta}, L_2(Q)\right)}du < \infty \qquad (6.11)$$

*or*

$$\sup_{\theta \in \Theta_n^\tau} \int_0^\infty \sqrt{\log N_{[\,]}\left(u\|F_{n,\theta}\|_{L_2(P)}, \mathcal{F}_{n,\theta}, L_2(P)\right)}du < \infty \qquad (6.12)$$

*Then, the sequence $\{Z_n(h,\hat{\theta}_n) : h \in \mathcal{H}\}$ is asymptotically tight in $l^\infty(\mathcal{H})$. Here, $N_{[\,]}()$ and $N()$ denote the bracketing and covering numbers respectively and the supremum in (6.11) is taken over all discrete probability measures $Q$.*

The measurability required for the class $\mathcal{F}_{n,\delta}$ is in the following sense. For any vector $\{e_1, \ldots, e_n\} \in \{-1, 1\}^n$, the map

$$(V_1, V_2, \ldots, V_n, \hat{\theta}, e_1, \ldots, e_n) \mapsto \sup_{g_{n,\hat{\theta}} \in \mathcal{F}_{n,\delta}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_i g_{n,\hat{\theta}}(V_i) \right| \tag{6.13}$$

is assumed to be jointly measurable. This is very much in the spirit of the $P$-measurability assumption made for Donsker results involving covering numbers (e.g., van der Vaart and Wellner (1996, Theorem 2.5.2)) and can be justified readily in many applications.

We prove the above result assuming (6.11). The proof follows the road map of that for Theorems 2.5.2 and 2.11.1 of van der Vaart and Wellner (1996) and is outlined in Section E.3 of Appendix E.

In our examples, the form of the limit process does not depend on the weak limit of the first stage estimates, and can be derived using the following lemma.

**Lemma 6.4.** *Consider the setup of Theorem 6.3. Additionally, assume that for any* $\tau > 0$*,*

1. *The covariance function*

$$C_n(h_1, h_2, \theta) = P f_{n,h_1,\theta} f_{n,h_2,\theta} - P f_{n,h_1,\theta} P f_{n,h_2,\theta}$$

   *converges pointwise to* $C(h_1, h_2)$ *on* $\mathcal{H} \times \mathcal{H}$*, uniformly in* $\theta$*,* $\theta \in \Theta_n^\tau$*.*

2. *The functions* $\zeta_n(h, \theta)$ *converges pointwise to a function* $\zeta(h)$ *on* $\mathcal{H}$*, uniformly in* $\theta$*,* $\theta \in \Theta_n^\tau$*.*

*Let* $Z(h)$ *be a Gaussian process with drift* $\zeta(\cdot)$ *and covariance kernel* $C(\cdot, \cdot)$*. Then, the process* $Z_n(\cdot, \hat{\theta}_n)$ *converges weakly to* $Z(\cdot)$ *in* $\ell^\infty(\mathcal{H})$*.*

We prove a stronger result where we allow for the limit distribution of the first stage estimates to affect the limit process $Z$. The proof is given in Section E.4 of Appendix E.

In our applications, the process $Z_n(h, \hat{\theta}_n)$ is defined for $h$ in a Euclidean space, say $\tilde{\mathcal{H}} = \mathbb{R}^p$ and Theorem 6.1 is used to show that $\hat{h}_n := s_n(\hat{d}_n - d_n)$, which assumes values in $\tilde{\mathcal{H}}$, is $O_p(1)$. The process $Z_n$ is viewed as living in $\mathcal{B}_{loc}(\mathbb{R}^p) = \{f : \mathbb{R}^p \mapsto \mathbb{R} : f$ is bounded on $[-L, L]^p$ for any $L > 0\}$, the space of locally bounded functions on $\mathbb{R}^p$.

To deduce the limit distribution of $\hat{h}_n$, we first show that for a Gaussian process $Z(h)$ in $C_{min}(\mathbb{R}^p) = \{f \in \mathcal{B}_{loc}(\mathbb{R}^p) : f$ possesses a unique minimum and $f(x) \to \infty$ as $\|x\| \to \infty\}$, the process $Z_n(h, \hat{\theta}_n)$ converges to $Z(h)$ in $\mathcal{B}_{loc}(\mathbb{R}^p)$. This is accomplished by showing that on every $[-L, L]^p$, $Z_n(h, \hat{\theta}_n)$ converges to $Z(h)$ on $\ell^\infty([L, L]^p)$, using Theorem 6.3 and Lemma 6.4. An application of the argmin continuous mapping theorem (Theorem 2.7) of Kim and Pollard (1990) now yields the desired result, i.e., $\hat{h}_n \xrightarrow{d} \text{argmin}_{h \in \mathbb{R}^p} Z(h)$.

Next, we summarize what has been discussed above to provide a generic approach to multi-stage problems.

*Rate of convergence.*

1. With $\hat{\theta}_n$ denoting the first stage estimate, identify the second stage criterion as a bivariate function $\mathbb{M}_n(d, \hat{\theta}_n)$ and its population equivalent $M_n(d, \hat{\theta}_n)$. A useful choice for $M_n$ is $M_n(d, \theta) = E\left[\mathbb{M}_n(d, \theta)\right]$. The non-random process $M_n$ is at its minimum at $d_n$ which either equals the parameter of interest $d_0$ or is asymptotically close to it.

2. Arrive at $\rho_n(d, d_n)$ using (6.2) which typically involves a second order Taylor expansion when $M_n$ is smooth. The distance $\rho_n$ is typically some function of the Euclidean metric.

3. Justify the convergence $P\left(\rho_n(\hat{d}_n, d_n) \geq \kappa_n\right)$ to zero using Lemma 6.2 if needed and derive a bound on the modulus of continuity as in (6.3). This typically requires VC or bracketing arguments such as Theorem 2.14.1 of van der Vaart and Wellner (1996). With suitably chosen $K_\tau$, $\Theta_n^\tau$ can be chosen to be shrinking

sets of type $[\theta_n - K_\tau/n^\nu, \theta_n + K_\tau/n^\nu]$, when a result of the type $n^\nu(\hat{\theta}_n - \theta_n) = O_p(1)$ holds. Such choices typically yield efficient bounds for (6.3).

4. Derive the rate of convergence using Theorem 6.1.

   *Limit Distribution.*

5. Express the local process $Z_n$ as an empirical process acting on a class of functions and a drift term (see (6.6)).

6. Use Theorem 6.3 and Lemma 6.4 to derive the limit process $Z$ and apply argmin continuous mapping to derive the limiting distribution of $\hat{d}_n$.

The following sections illustrate applications of the above results.

## 6.2 Change-point model with fainting signal

We consider a change-point model of the form $Y = m_n(X) + \epsilon$, where

$$m_n(x) = \alpha_n 1[x \leq d_0] + \beta_n 1[x \geq d_0]$$

for an unknown $d_0 \in (0, 1)$ and $\beta_n - \alpha_n = c_0 n^{-\xi}$, $c_0 > 0$ and $\xi < 1/2$. The errors $\epsilon$ are independent of $X$ and have mean 0 and variance $\sigma^2$. In contrast with the change-point model considered in Lan et al. (2009), the signal in the model $\beta_n - \alpha_n$ decreases with $n$. A similar model with decreasing signal was studied in Müller and Song (1997). We assume that the experimenter has the freedom to choose the design points to sample from but has a fixed budget $n$. We apply the following two-stage approach.

1. At stage one, sample $n_1 = pn$ covariate values, ($p \in (0, 1)$), from a uniform design on $\mathcal{D} = [0, 1]$ and, from the obtained data, $(Y_i^{(1)}, X_i^{(1)})_{i=1}^{n_1}$, estimate $\alpha_n$,

$\beta_n$ and $d_0$ by

$$\hat{\theta}_{n_1} = \left(\hat{\alpha}, \hat{\beta}, \hat{d}_1\right)$$
$$= \underset{\alpha,\beta,d}{\operatorname{argmin}} \sum_{i=1}^{n_1} \left[(Y_i^{(1)} - \alpha)^2 1\left[X_i^{(1)} \leq d\right] + (Y_i^{(1)} - \beta)^2 1\left[X_i^{(1)} > d\right]\right].$$

These are simply the least squares estimates.

2. For $K > 0$ and $\gamma > 0$, sample the remaining $n_2 = (1-p)n$ covariate-response pairs $\{Y_i^{(2)}, X_i^{(2)}\}$, where

$$Y_i^{(2)} = \alpha_n 1[X_i^{(2)} \leq d_0] + \beta_n 1[X_i^{(2)} \geq d_0] + \epsilon_i$$

and $X_i$'s are sampled uniformly from the interval $\mathcal{D}_{\hat{\theta}_{n_1}} = [\hat{d}_1 - K{n_1}^{-\gamma}, \hat{d}_1 - K{n_1}^{-\gamma}]$. Obtain an updated estimate of $d_0$ by

$$\hat{d}_2 = \underset{d \in \mathcal{D}_{\hat{\theta}_n}}{\operatorname{argmin}} \sum_{i=1}^{n_2} \left[(Y_i^{(2)} - \hat{\alpha})^2 1\left[X_i^{(2)} \leq d\right] + (Y_i^{(2)} - \hat{\beta})^2 1\left[X_i^{(2)} > d\right]\right]. \quad (6.14)$$

Here, $\gamma$ is chosen such that $P\left(d_0 \in [\hat{d}_1 - K{n_1}^{-\gamma}, \hat{d}_1 - K{n_1}^{-\gamma}]\right)$ converges to 1. Intuitively, this condition compels the second stage design interval to contain $d_0$ with high probability. This is needed as the objective function relies on the dichotomous behavior of the regression function on either side of $d_0$ for estimating the change-point. If the second stage interval does not include $d_0$ (with high probability), the stretch of the regression function $m_n$ observed (with noise) is simply flat, thus failing to provide information about $d_0$.

In Bhattacharya and Brockwell (1976) and Bhattacharya (1987), similar models were studied in a one-stage fixed design setting. By a minor extension of their results, it can be shown that ${n_1}^{\nu}(\hat{d}_1 - d_0) = O_p(1)$ for $\nu = 1 - 2\xi$, $\sqrt{n_1}(\hat{\alpha} - \alpha_n) = O_p(1)$ and $\sqrt{n_1}(\hat{\beta} - \beta_n) = O_p(1)$. Hence, any choice of $\gamma < \nu$ suffices.

For simplicity, we assume that the experimenter works with a uniform random design at both stages. An extension to designs with absolutely continuous positive densities supported on an interval is straightforward.

The expression in (6.14) can be simplified to yield

$$\hat{d}_2 = \operatorname*{argmin}_{d \in \mathcal{D}_{\hat{\theta}_{n_1}}} \mathbb{M}_{n_2}(d, \hat{\theta}) \tag{6.15}$$

where for $\theta = (\alpha, \beta, \mu) \in \mathbb{R}^3$,

$$\mathbb{M}_{n_2}(d, \theta) \quad = \quad \frac{\operatorname{sgn}(\beta - \alpha)}{n_2} \sum_{i=1}^{n_2} \left( Y_i^{(2)} - \frac{\alpha + \beta}{2} \right) \left( 1\left[ X_i^{(2)} \leq d \right] - 1\left[ X_i^{(2)} \leq d_0 \right] \right)$$

with $X_i \sim \text{Uniform}[\mu - K{n_1}^{-\gamma}, \mu + K{n_1}^{-\gamma}]$, $\hat{\theta}_n = (\hat{\alpha}, \hat{\beta}, \hat{d}_1)$ and sgn denoting the sign function. We take $M_{n_2}(d, \theta) = E\left[\mathbb{M}_{n_2}(d, \theta)\right]$ to apply Theorem 6.1, which yields the following result on the rate of convergence of $\hat{d}_2$.

**Theorem 6.5.** *For $\hat{d}_2$ defined in* (6.15) *and $\eta = 1 + \gamma - 2\xi$*

$$n^{\eta}(\hat{d}_2 - d_0) = O_p(1).$$

*Proof.* As $n_1$, $n_2$ and $n$ are of the same order, we deduce bounds in terms of $n$ only. For notational ease, we first consider the situation where $d \geq d_0$. Recall that $\theta = (\alpha, \beta, \mu)$. Also, let

$$\begin{aligned} \Theta_{n_1}^{\tau} = & \left[ \alpha_n - \frac{K_\tau}{\sqrt{n_1}}, \alpha_n + \frac{K_\tau}{\sqrt{n_1}} \right] \times \left[ \beta_n - \frac{K_\tau}{\sqrt{n_1}}, \beta_n + \frac{K_\tau}{\sqrt{n_1}} \right] \times \\ & \left[ d_0 - \frac{K_\tau}{{n_1}^{\nu}}, d_0 + \frac{K_\tau}{{n_1}^{\nu}} \right], \end{aligned} \tag{6.16}$$

where $K_\tau$ is chosen such that $P\left(\hat{\theta}_{n_1} \in \Theta_{n_1}^{\tau}\right) > 1 - \tau$. For $\theta \in \Theta_{n_1}^{\tau}$, $\beta - \alpha \geq c_0 n^{-\xi} - 2K_\tau/\sqrt{n_1}$. As $\xi < 1/2$, $\operatorname{sgn}(\beta - \alpha) = 1$ for $n > N_\tau^{(1)} := (2K_\tau/(\sqrt{p}c_0))^{2/(2-\xi)}$.

Also, for $x > d_0$, $m_n(x) = \beta_n$ and thus,

$$\mathbb{M}_{n_2}(d,\theta) \;=\; \mathbb{P}_{n_2}\left[g_{n_2,d,\theta}(V)\right]$$

where for $V = (U,\epsilon)$, $U \sim$ Uniform$[-1,1]$,

$$\begin{aligned} g_{n_2,d,\theta}(V) &= \left(\beta_n + \epsilon - \frac{\beta+\alpha}{2}\right) 1\left[\mu + K n_1^{-\gamma} U \in (d_0, d]\right] \\ &= \left(\beta_n + \epsilon - \frac{\beta+\alpha}{2}\right) 1\left[U \in \left(\frac{d_0-\mu}{K n_1^{-\gamma}}, \frac{d-\mu}{K n_1^{-\gamma}}\right]\right]. \end{aligned}$$

Consequently, for $n > N_\tau^{(1)}$,

$$M_{n_2}(d,\theta) \;=\; \frac{1}{2}\left(\beta_n - \frac{\beta+\alpha}{2}\right)\lambda\left([-1,1] \cap \left(\frac{d_0-\mu}{K n_1^{-\gamma}}, \frac{d-\mu}{K n_1^{-\gamma}}\right]\right).$$

As $\gamma < \nu$, $d_0 \in \mathcal{D}_\theta$ for all $\theta \in \Theta_{n_1}^\tau$, for $n > N_\tau^{(2)} := (1/p)(K_\tau/K)^{1/(\nu-\gamma)}$ and hence, the intervals

$$\left\{\left((d_0-\mu)/(K n_1^{-\gamma}), (d-\mu)/(K n_1^{-\gamma})\right] : d > d_0, d \in \mathcal{D}_\theta, \theta \in \Theta_{n_1}^\tau\right\}$$

are all contained in $[-1,1]$ for large $n$. Therefore, for $n > N_\tau^{(3)} := \max(2N_\tau^{(1)}, N_\tau^{(2)})$,

$$M_{n_2}(d,\theta) \;=\; \frac{1}{2}\left(\beta_n - \frac{\beta+\alpha}{2}\right)\frac{d-d_0}{K n_1^{-\gamma}}.$$

Note that $M_{n_2}(d_0,\theta) = 0$ for all $\theta \in \mathbb{R}^3$. Further, let $\rho_n^2(d,d_0) = n^{\gamma-\xi}|d-d_0|$. Then, for $n > N_\tau^{(3)}$,

$$\begin{aligned} M_{n_2}(d,\theta) - M_{n_2}(d_0,\theta) &\geq \left(\beta_n - \frac{\beta_n+\alpha_n}{2} - \frac{K_\tau}{\sqrt{n_1}}\right)\frac{d-d_0}{2K n_1^{-\gamma}} \\ &= \left(\frac{\beta_n-\alpha_n}{2} - \frac{K_\tau}{\sqrt{n_1}}\right)\frac{d-d_0}{2K n_1^{-\gamma}} \end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{c_0 n^{-\xi}}{2} - \frac{K_\tau}{\sqrt{n_1}}\right)\frac{d-d_0}{2Kn_1^{-\gamma}} \\
&\geq c_\tau \rho_n^2(d,d_0),
\end{aligned} \tag{6.17}
$$

for some $c_\tau > 0$ (depending on $\tau$ through $K_\tau$). The last step follows from the fact that $\xi < 1/2$. Also, the above lower bound can be shown to hold for the case $d > d_0$ as well. Further, to apply Theorem 6.1, we need to bound

$$
\sup_{\theta\in\Theta_{n_1}^\tau} E^* \sup_{\substack{|d-d_0|<n^{\xi-\gamma}\delta^2,\\ d\in\mathcal{D}_\theta}} \sqrt{n_2}\left|(\mathbb{M}_{n_2}(d,\theta) - M_{n_2}(d,\theta)) - (\mathbb{M}_{n_2}(d_0,\theta) - M_{n_2}(d_0,\theta))\right|. \tag{6.18}
$$

Note that for $d > d_0$, the expression in $|\cdot|$ equals $(1/\sqrt{n_2})\mathbb{G}_{n_2} g_{n_2,d,\theta}$. The class of functions $\mathcal{F}_{\delta,\theta} = \{g_{n_2,d,\theta} : 0 \leq d - d_0 < n^{\xi-\gamma}\delta^2, d \in \mathcal{D}_\theta\}$ is VC with index at most 3 and is enveloped by

$$
M_{\delta,\theta}(V) = \left(|\epsilon| + \frac{\beta_n - \alpha_n}{2} + \frac{K_\tau}{\sqrt{n_1}}\right) 1\left[U \in \left[\frac{d_0-\mu}{Kn_1{}^{-\gamma}}, \frac{d_0 - \mu + \delta^2 n^{\xi-\gamma}}{Kn_1{}^{-\gamma}}\right]\right].
$$

Note that

$$
\begin{aligned}
&E\left[M_{\delta,\theta}(V)\right]^2 \\
&= \frac{1}{2}E\left[\left(|\epsilon| + \frac{\beta_n - \alpha_n}{2} + \frac{K_\tau}{\sqrt{n_1}}\right)^2\right]\lambda\left[[-1,1]\cap\left[\frac{d_0-\mu}{Kn_1{}^{-\gamma}}, \frac{d_0-\mu+\delta^2 n^{\xi-\gamma}}{Kn_1{}^{-\gamma}}\right]\right] \\
&\leq \frac{1}{2}E\left[\left(|\epsilon| + \frac{\beta_n - \alpha_n}{2} + \frac{K_\tau}{\sqrt{n_1}}\right)^2\right]\lambda\left[\frac{d_0-\mu}{Kn_1{}^{-\gamma}}, \frac{d_0-\mu+\delta^2 n^{\xi-\gamma}}{Kn_1{}^{-\gamma}}\right] \\
&\leq C_\tau^2 \frac{n^{\xi-\gamma}\delta^2}{n^{-\gamma}} = C_\tau^2 n^\xi \delta^2,
\end{aligned}
$$

where $C_\tau$ is positive constant (it depends on $\tau$ through $K_\tau$). Further, the uniform entropy integral for $\mathcal{F}_{\delta,\theta}$ is bounded by a constant which only depends upon the

VC-indices, i.e., the quantity

$$J(1, \mathcal{F}_{\delta,\theta}) = \sup_{Q} \int_{0}^{1} \sqrt{1 + \log N(u\|M_{\delta,\theta}\|_{Q,2}, \mathcal{F}_{\delta,\theta}, L_2(Q))}du$$

is bounded, where $N(\cdot)$ denotes the covering number; see Theorems 9.3 and 9.15 of Kosorok (2008) for more details. Using Theorem 2.14.1 of van der Vaart and Wellner (1996),

$$E^* \sup_{\substack{0 \le d - d_0 < n^{\xi-\gamma}\delta^2 \\ d \in \mathcal{D}_\theta}} |\mathbb{G}_{n_2} g_{n_2,d,\theta}| \le J(1, \mathcal{F}_{\delta,\tau})\|M_{\delta,\theta}\|_2 \le C_\tau n^{\xi/2}\delta. \tag{6.19}$$

Note that this bound does not depend on $\theta$ and can be shown to hold for the case $d \le d_0$ as well. Hence, we get the bound $\phi_n(\delta) = n^{\xi/2}\delta$ on the modulus of continuity. For $n > N_\tau^{(3)}$, (6.17) holds for all $d \in \mathcal{D}_\theta$, and (6.19) is valid for all $\delta > 0$. Hence, we do not need to justify consistency with respect to $\rho_n$. For $r_n = n^{1/2-\xi/2}$, the relation $r_n^2\phi_n(1/r_n) \le \sqrt{n}$ is satisfied. Consequently, $r_n^2(n^{\gamma-\xi}(\hat{d}_n - d_0)) = n_2^\eta(\hat{d}_n - d_0) = O_p(1)$. □

To deduce the limit distribution of $\hat{d}_2$, consider the process

$$Z_{n_2}(h, \theta) = \frac{1}{n_2^\xi} \sum_{i=1}^{n_2} \left(Y_i^{(2)} - \frac{\alpha + \beta}{2}\right) \left(1\left[X_i^{(2)} \le d_0 + hn^{-\eta}\right] - 1\left[X_i^{(2)} \le d_0\right]\right) \tag{6.20}$$

with $X_i^{(2)} \sim \text{Uniform}[\mu - Kn_1{}^{-\gamma}, \mu + Kn_1{}^{-\gamma}]$. Note that

$$n^\eta(\hat{d}_2 - d_0) = \operatorname*{argmin}_h Z_{n_2}(h, \hat{\theta}).$$

It is convenient to write $Z_{n_2}$ as

$$Z_{n_2}(h, \theta) = \mathbb{G}_{n_2} f_{n_2,h,\theta}(V) + \zeta_{n_2}(h, \theta), \tag{6.21}$$

where $\zeta_{n_2}(h,\theta) = \sqrt{n_2} P f_{n_2,h,\theta}(V)$ and

$$\begin{aligned} f_{n_2,h,\theta}(V) &= n_2^{1/2-\xi}\left(m_n(\mu + UKn_1^{-\gamma}) + \epsilon - \frac{\alpha+\beta}{2}\right) \times \\ &\quad \left(1\left[\mu + UKn_1^{-\gamma} \le d_0 + hn^{-\eta}\right] - 1\left[\mu + UKn_1^{-\gamma} \le d_0\right]\right). \end{aligned}$$

This is precisely the form of the local process needed for Theorem 6.3. We next use it to deduce the weak limit of the process $Z_{n_2}(h,\hat{\theta})$.

**Theorem 6.6.** *Let $B$ be a standard Brownian motion on $\mathbb{R}$ and*

$$Z(h) = \sqrt{\frac{(1-p)^{1-2\xi}p^{\gamma}}{2K}}\sigma B(h) + \frac{(1-p)^{1-\xi}p^{\gamma}}{2K}\frac{c_0}{2}|h|.$$

*Then, the sequence of stochastic process $Z_{n_2}(h)$, $h \in \mathbb{R}$ are asymptotically tight and converge weakly to $Z(t)$.*

*Proof.* For any $L > 0$, we start by justifying the conditions of Theorem 6.3 to prove tightness of the process $Z_{n_2}(h,\hat{\theta}_{n_1})$, for $h \in [-L, L]$. For sufficiently large $n$, the set $\{h : d_0 + h/n^{\eta} \in \mathcal{D}_{\theta}\}$ contains $[-L, L]$ for all $\theta \in \Theta_{n_1}^{\tau}$ and hence, it is not necessary to extend $Z_{n_2}$ (equivalently, $f_{n_2,h,\theta}$) as done in (6.5). Further, for a fixed $\theta \in \Theta_{n_1}^{\tau}$ (defined in (6.16)), an envelope for the class of functions $\{f_{n_2,h,\theta} : |h| \le L\}$ is given by

$$\begin{aligned} F_{n_2,\theta}(V) &= n_2^{1/2-\xi}\left(\frac{\beta_n - \alpha_n}{2} + \frac{K_{\tau}}{\sqrt{n_1}} + |\epsilon|\right) \times \\ &\quad 1\left[\mu + UKn_1^{-\gamma} \in [d_0 - Ln^{-\eta}, d_0 + Ln^{-\eta}]\right]. \end{aligned}$$

Note that

$$PF_{n_2,\theta}^2 \lesssim n^{1-2\xi}\left(\left(\frac{\beta_n - \alpha_n}{2} + \frac{K_{\tau}}{\sqrt{n_1}}\right)^2 + \sigma^2\right)\frac{2Ln^{-\eta}}{2Kn_1^{-\gamma}}$$

As $\eta = 1+\gamma-2\xi$, the right hand side is $O(1)$. Moreover, the bound is uniform in $\theta$, $\theta \in \Theta_{n_1}^{\tau}$. Let $K_0$ be a constant (depending on $\tau$) such that $K_0 \ge (\beta_n - \alpha_n)/2 + K_{\tau}/\sqrt{n_1}$.

Then, for $t > 0$, $PF^2_{n_2,\theta} 1[F_{n_2,\theta} > \sqrt{n_2}t]$ is bounded by

$$n^{1-2\xi} P\left((K_0 + |\epsilon|)^2 1\left[\mu + UKn_1^{-\gamma} \in [d_0 - Ln^{-\eta}, d_0 + Ln^{-\eta}]\right] \times \right.$$
$$\left. 1\left[n^{1/2-\xi}(K_0 + |\epsilon|) > \sqrt{n_2}t\right]\right)$$

As $\epsilon$ and $U$ are independent, the above is bounded up to a constant by

$$P(K_0 + |\epsilon|)^2 1\left[(K_0 + |\epsilon|) > \sqrt{p}n^{\xi}t\right]$$

which goes to zero. This justifies condition (6.7) and (6.8) of Theorem 6.3. Let $\tilde{\rho}(h_1, h_2) = |h_1 - h_2|$. For any $L > 0$, the space $[-L, L]$ is totally bounded with respect to $\tilde{\rho}$. For $h_1, h_2 \in [-L, L]$ and $\theta \in \Theta^{\tau}_{n_1}$, we have

$$P(f_{n_2,h_1,\theta} - f_{n_2,h_2,\theta})^2 \lesssim n^{1-2\xi} \frac{|h_1 - h_2| n^{-\eta}}{2Kn_1^{-\gamma}} E\left[K_0 + |\epsilon|\right]^2.$$

The right side is bounded up to a constant by $|h_1 - h_2|$ for all choices of $\theta$, $\theta \in \Theta^{\tau}_{n_1}$. Hence, condition (6.9) is satisfied as well. Condition (6.10) can be justified in a manner mentioned later. Further, the class of functions $\{f_{n_2,h,\theta} : |h| \leq L\}$ is VC of index at most 3 with envelope $F_{n_2,\theta}$. Hence, it has a bounded entropy integral with the bound only depending on the VC index of the class (see Theorems 9.3 and 9.15 of Kosorok (2008)) and hence, condition (6.11) is also satisfied. Also, the measurability condition (6.13) can be shown to hold by approximating $\mathcal{F}_{n_2,\delta} = \{f_{n_2,h_1,\theta} - f_{n_2,h_2,\theta} : |h_1 - h_2| < \delta\}$ (defined in Theorem 6.3) by the countable class involving only rational choices of $h_1$ and $h_2$. Note that the supremum over this countable class is measurable and it agrees with supremum over $\mathcal{F}_{n_2,\delta}$. Thus $\mathbb{G}_{n_2} f_{n_2,h,\hat{\theta}}$ is tight in $l^{\infty}([-L, L])$.

Next, we apply Lemma 6.4 to deduce the limit process. Note that for $\theta \in \Theta_{n_1}^{\tau}$ and $|h| \leq L$,

$$\begin{aligned}
\zeta_{n_2}(h,\theta) &= n_2^{1-\xi}\left(\alpha_n 1(h \leq 0) + \beta_n 1(h > 0) - \frac{\alpha+\beta}{2}\right)\frac{hn^{-\eta}}{2Kn_1^{-\gamma}} \\
&= (1-p)^{1-\xi}\left(\alpha_n 1(h \leq 0) + \beta_n 1(h > 0) - \frac{\alpha+\beta}{2}\right)\frac{hn^{\xi}}{2Kp^{-\gamma}} \\
&= \frac{(1-p)^{1-\xi}p^{\gamma}n^{\xi}}{2K}h\left(\alpha_n 1(h \leq 0) - \beta_n 1(h > 0) + \frac{\alpha_n+\beta_n}{2}\right) + R_n.
\end{aligned}$$

The remainder term $R_n$ in the last step accounts for replacing $\alpha+\beta$ by $\alpha_n+\beta_n$ in the expression for $\zeta_{n_2}$ and is bounded (uniformly in $\theta \in \Theta_{n_1}^{\tau}$) up to a constant by

$$n^{\xi}L\left(|\alpha_n - \alpha| + |\beta_n - \beta|\right) = O(n^{\xi-1/2}).$$

As $\xi < 1/2$, $\sqrt{n_2}Pf_{n_2,h,\theta}$ converges uniformly to $|h|\left((1-p)^{1-\xi}p^{\gamma}c_0\right)/(4K)$. Condition (6.10) can be justified by calculations parallel to the above. Further, $Pf_{n_2,h,\theta} = \zeta_{n_2}(h,\theta)/\sqrt{n_2}$ converges to zero (uniformly) and hence, the covariance function of the limiting Gaussian process (for $h_1, h_2 > 0$) is given by

$$\begin{aligned}
&\lim_{n\to\infty} Pf_{n_2,h_1,\theta}f_{n_2,h_1,\theta} \\
&\quad = \lim_{n\to\infty} n_2^{1-2\xi}\left[\left(\alpha_n 1(h \leq 0) + \beta_n 1(h > 0) - \frac{\alpha+\beta}{2}\right)^2 + \sigma^2\right]\frac{h_1 \wedge h_2 n^{-\eta}}{2Kn_1^{-\gamma}} \\
&\quad = \frac{(1-p)^{1-2\xi}p^{\gamma}\sigma^2}{2K}(h_1 \wedge h_2).
\end{aligned}$$

Analogous results can be established for other choices of $(h_1, h_2) \in [-L, L]^2$. Also, the above convergence can be shown to be uniform in $\theta$ by a calculation similar to that done for $\zeta_{n_2}$. This justifies the form of the limit $Z$. Hence, we get the result. □

*Comparison with results from van der Vaart and Wellner (2007).* As mentioned earlier, van der Vaart and Wellner (2007) derived sufficient conditions to prove results of the form $\sup_{d\in\mathcal{D}} \left|\mathbb{G}_n(f_{d,\hat{\theta}} - f_{d,\theta_0})\right| \xrightarrow{p} 0$, where $\{f_{d,\theta} : d \in \mathcal{D}, \theta \in \Theta\}$ is a suitable class of measurable functions and $\hat{\theta}$ is a consistent estimate of $\theta_0$. If such a result were to hold in the above model, the derivation of the limit process would boil down to working with the process $\{\mathbb{G}_n f_{d,\theta_0} : d \in \mathcal{D}\}$, which is much simpler to work with. However, we show below that for $h \neq 0$,

$$T_{n_2} := (Z_{n_2}(h, \alpha_n, \beta_n, \hat{d}_1) - Z_{n_2}(h, \alpha_n, \beta_n, d_0)) \tag{6.22}$$

*does not converge in probability to zero,* let alone the supremum of the above over $h$ in compact sets and hence, the results in van der Vaart and Wellner (2007) do not apply.

**Proposition 6.7.** *Let $\pi_0^2 := \sigma^2 p^\gamma (1-p)^{1-2\xi}|h|/K$ and $T_{n_2}$ be as defined in (6.22). Then, for $h \neq 0$, $T_{n_2}$ converges to a normal distribution with mean 0 and variance $\pi_0^2$.*

The proof is given in Section E.5 of Appendix E. We now provide the limiting distribution of $\hat{d}_2$.

**Theorem 6.8.** *The process $Z$ possesses a unique tight argmin almost surely and for $\lambda_0 = (8K\sigma^2)/(c_0^2(1-p)p^\gamma)$,*

$$n^\eta(\hat{d}_2 - d_0) \xrightarrow{d} \operatorname*{argmin}_h Z(h) \stackrel{d}{=} \lambda_0 \operatorname*{argmin}_v \left[B(v) + |v|\right].$$

**Remark 6.9.** *We considered a uniform random design for sampling at both stages. The results extend readily to other suitable designs. For example, if the second stage design points are sampled as $X_i^{(2)} = \hat{d}_1 + V_i K n_1^{-\gamma}$, where $V_i$'s are i.i.d. realizations from a distribution with a positive continuous density $\psi$ supported on a compact set containing an interval around zero, it can be shown that $\hat{d}_2$ attains the same rate of convergence. The limit distribution has the same form as above with $\lambda_0$ replaced by*

$\lambda_0/\psi(d_0)$.

*Proof.* As $\mathrm{Var}(Z(t) - Z(s)) \neq 0$, uniqueness of the argmin follows immediately from Lemma 2.6 of Kim and Pollard (1990). Also, $Z(h) \to \infty$ as $|h| \to \infty$ almost surely. This is true as $Z(h) = |h|\,[\sigma^2 B(h)/|h| + c/2]$ with $B(h)/|h|$ converging to zero almost surely as $|h| \to \infty$. Consequently, $Z \in C_{min}(\mathbb{R})$ with probability one and the unique argmin of $Z$ is tight. An application of argmin continuous mapping theorem (Kim and Pollard, 1990, Theorem 2.7) then gives us distributional convergence. By dropping a constant multiple, it can be seen that

$$\operatorname*{argmin}_h Z(h) = \operatorname*{argmin}_h \left[\sigma B(h) + \sqrt{\frac{(1-p)p^\gamma}{2K}}\frac{c_0}{2}|h|\right].$$

As $\sigma\sqrt{\lambda_0} = \sqrt{((1-p)p^\gamma)/(2K)}(c_0\lambda_0)/2$, by the rescaling property of Brownian motion,

$$\begin{aligned}
&\operatorname*{argmin}_h \left[\sigma B(h) + \sqrt{\frac{(1-p)p^\gamma}{2K}}\frac{c_0}{2}|h|\right] \\
&\qquad = \lambda_0 \operatorname*{argmin}_v \left[\sigma B(\lambda_0 v) + \sqrt{\frac{(1-p)p^\gamma}{2K}}\frac{c_0}{2}|\lambda_0||v|\right] \\
&\qquad \stackrel{d}{=} \lambda_0 \operatorname*{argmin}_v \left[\sigma\sqrt{\lambda_0}B(v) + \sqrt{\frac{(1-p)p^\gamma}{2K}}\frac{c_0}{2}\lambda_0|v|\right] \\
&\qquad = \lambda_0 \operatorname*{argmin}_v \left[B(v) + |v|\right].
\end{aligned}$$

The result follows. ☐

*Optimal allocation.* The interval from which the covariates are sampled at the second stage is chosen such that the change-point $d_0$ would be contained in the prescribed interval with high probability, i.e., we pick $K$ and $\gamma$ such that $P\left(d_0 \in [\hat{d}_1 - K{n_1}^{-\gamma}, \hat{d}_1 - K{n_1}^{-\gamma}]\right)$ converges to 1. But, in practice for a fixed $n$, a suitable choice would be

$$K{n_1}^{-\gamma} \approx \frac{C_{\tau/2}}{n_1^{1-2\xi}}$$

with $C_{\tau/2}$ being the $(1-\tau/2)$th quantile of the limiting distribution of $n_1^{1-2\xi}(\hat{d}_1 - d_0)$ which is symmetric around zero. As $\mathrm{argmin}_v\,[B(v) + |v|]$ is a symmetric random variable, the variance of $(\hat{d}_2 - d_0)$ would then be (approximately) smallest when

$$\begin{aligned}\frac{\lambda_0}{n^\eta} &= \frac{8K\sigma^2}{c_0^2(1-p)p^\gamma n^\eta} = \frac{8\sigma^2 C_{\tau/2}}{c_0^2(1-p)p^\gamma n^\eta n_1^{1-\gamma-2\xi}} \\ &= \frac{8\sigma^2 C_{\tau/2}}{c_0^2(1-p)p^{1-2\xi}n^{2(1-2\xi)}}\end{aligned}$$

is at its minimum. This yields the optimal choice of $p$ to be $p_{opt} = (1-2\xi)/(2(1-\xi))$.

## 6.3 Inverse isotonic regression

In this section, we consider the problem of estimating the *inverse* of a monotone regression function at a pre-specified point $t_0$ using multi-stage procedures. Responses $(Y, X)$ are obtained from a model of the form $Y = r(X) + \epsilon$, where $r$ is a monotone function on [0,1] and the experimenter has the freedom to choose the design points. It is of interest to estimate the threshold $d_0 = r^{-1}(t_0)$ for some $t_0$ in the interior of the range of $r$ with $r'(d_0) > 0$. The estimation procedure is summarized below:

1. At stage one, sample $n_1 = p \times n$ covariate values uniformly from $[0, 1]$ and, from the data, $(Y_i^{(1)}, X_i^{(1)})_{i=1}^{n_1}$, obtain the isotonic regression estimate $\hat{r}_{n_1}$ of $r$ (see Robertson et al. (1988, Chapter 1)) and, subsequently, an estimate $\hat{d}_1 = \hat{r}_{n_1}^{-1}(t_0)$ of $d_0$.

2. For $K > 0$ and $\gamma > 0$, sample the remaining $n_2 = (1-p)n$ covariate-response pairs $\{Y_i^{(2)}, X_i^{(2)}\}_{i=1}^{n_2}$, where

$$Y_i^{(2)} = r(X_i^{(2)}) + \epsilon_i, \quad X_i^{(2)} \sim \text{Uniform}[\hat{d}_1 - Kn_1^{-\gamma}, \hat{d}_1 + Kn_1^{-\gamma}].$$

   Obtain an updated estimate $\hat{d}_2 = \hat{r}_{n_2}^{-1}(d_0)$ of $d_0$, where $\hat{r}_{n_2}$ is the isotonic regression estimate based on $\{Y_i^{(2)}, X_i^{(2)}\}_{i \leq n_2}$. Also $\hat{r}_{n_2}^{-1}$ is the right continuous inverse of $\hat{r}_{n_2}$.

This procedure has been empirically studied in Tang et al. (2013). Here, we rigorously establish the limiting properties of $\hat{d}_2$. The parameter $\gamma$ is chosen such that $P\left(d_0 \in [\hat{d}_1 - Kn_1^{-\gamma}, \hat{d}_1 + Kn_1^{-\gamma}]\right)$ converges to 1. As $n^{1/3}(\hat{d}_1 - d_0) = O_p(1)$ (see, for example, Tang et al. (2011, Theorem 2.1)), any choice of $\gamma < 1/3$ suffices.

The *switching relationship* (Groeneboom, 1985, 1989) is useful in studying the limiting behavior of $\hat{r}_{n_2}$ through M-estimation theory. It simply relates the estimator $\hat{r}_{n_2}$ to the minima of a tractable process as follows. Let

$$V^0(x) = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i^{(2)} 1\left[X_i^{(2)} \leq x\right] \text{ and } G^0(x) = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i^{(2)} 1\left[X_i^{(2)} \leq x\right].$$

For $\hat{\theta}_{n_1} = \hat{d}_1$ and any $d \in [\hat{\theta}_{n_1} - Kn_1^{-\gamma}, \hat{\theta}_{n_1} - Kn_1^{-\gamma}]$, the following (switching) relation holds with probability one:

$$\hat{r}_{n_2}(d) \leq t \Leftrightarrow \underset{x \in [\hat{\theta}_{n_1} - Kn_1^{-\gamma}, \hat{\theta}_{n_1} - Kn_1^{-\gamma}]}{\operatorname{argmin}} \{V^0(x) - tG^0(x)\} \geq X_{(d)}^{(2)}, \tag{6.23}$$

where $X_{(d)}^{(2)}$ is the last covariate value $X_i^{(2)}$ to the left of $d$ and the argmin denotes the smallest minimizer if there are several. As $\hat{r}_{n_2}^{-1}$ is the right continuous inverse of $\hat{r}_{n_2}$, $\hat{r}_{n_2}(d) \leq t \Leftrightarrow d \leq \hat{r}_{n_2}^{-1}(t)$ and hence, using (6.23) at $t = t_0 = r(d_0)$, we get

$$\hat{d}_2 = \hat{r}_{n_2}^{-1}(t_0) \geq d \Leftrightarrow \underset{x \in [\hat{\theta}_{n_1} - Kn_1^{-\gamma}, \hat{\theta}_{n_1} - Kn_1^{-\gamma}]}{\operatorname{argmin}} \{V^0(x) - r(d_0)G^0(x)\} \geq X_{(d)}^{(2)}. \tag{6.24}$$

Let

$$\hat{x} = \underset{x \in [\hat{\theta}_{n_1} - Kn_1^{-\gamma}, \hat{\theta}_{n_1} - Kn_1^{-\gamma}]}{\operatorname{argmin}} \{V^0(x) - r(d_0)G^0(x)\}.$$

Note that both $\hat{x}$ and $\hat{d}_2$ are order statistics of $X$ ($\hat{r}_{n_2}(\cdot)$ and $V^0(\cdot) - r(d_0)G^0(\cdot)$ are piecewise constant functions). In fact, it can be shown using (6.24) twice (once at

$d = \hat{d}_2$ and the second time with $d$ being the order statistic to the immediate right of $\hat{d}_2$) that they are consecutive order statistics with probability one. Hence,

$$\hat{d}_2 = \hat{x} + O_p\left((2Kn_1^{-\gamma})\frac{\log n_2}{n_2}\right) = \hat{x} + O_p\left(\frac{\log n}{n^{1+\gamma}}\right). \tag{6.25}$$

The $O_p$ term in the above display corresponds to the order of the maximum of the differences between consecutive order statistics (from $n_2$ realizations from a uniform distribution on an interval of length $2Kn_1^{-\gamma}$). We will later show that $n^{(1+\gamma)/3}(\hat{x}-d_0) = O_p(1)$. As $n^{(1+\gamma)/3} = o(n^{1+\gamma}/\log n)$, it suffices to study the limiting behavior of $\hat{x}$ to arrive at the asymptotic distribution of $\hat{d}_2$. To this end, we start with an investigation of a version of the process $\{V^0(x) - r(d_0)G^0(x)\}$ at the resolution of the second stage "zoomed-in" neighborhood, given by

$$\mathbb{V}_{n_2}(u) = \mathbb{P}_{n_2}(Y^{(2)} - r(d_0))1\left[X^{(2)} \leq d_0 + un_2^{-\gamma}\right].$$

For $\mathcal{D}_{\hat{\theta}_{n_1}} = \left[n_2^{\gamma}(\hat{\theta}_{n_1} - Kn_1^{-\gamma}), n_2^{\gamma}(\hat{\theta}_{n_1} + Kn_1^{-\gamma})\right]$,

$$\hat{u} := n_2^{\gamma}(\hat{x} - d_0) = \underset{u \in \mathcal{D}_{\hat{\theta}_{n_1}}}{\operatorname{argmin}}\, \mathbb{V}_{n_2}(u).$$

Further, let $U \sim \text{Uniform}[-1,1]$ and $V = (U, \epsilon)$. Note that $X^{(2)} = \hat{\theta}_{n_1} + UKn_1^{-\gamma}$ and $Y^{(2)} = r(\hat{\theta}_{n_1} + UKn_1^{-\gamma}) + \epsilon$. Let

$$\begin{aligned} g_{n_2,u,\theta}(V) \quad &= \quad n_2^{\gamma}\left(r(\theta + UKn_1^{-\gamma}) + \epsilon - r(d_0)\right) \times \\ &\qquad \left(1\left[\theta + UKn_1^{-\gamma} \leq d_0 + un_2^{-\gamma}\right] - 1\left[\theta + UKn_1^{-\gamma} \leq d_0\right]\right). \end{aligned}$$

Also, let

$$\mathbb{M}_{n_2}(u, \theta) = \mathbb{P}_{n_2}\left[g_{n_2,u,\theta}(V)\right].$$

Then, $\hat{u} = \text{argmin}_{u \in \mathcal{D}_{\hat{\theta}_{n_1}}} \mathbb{M}_{n_2}\left(u, \hat{\theta}_{n_1}\right)$. Let $M_{n_2}(u, \theta) = P g_{n_2,u,\theta}$ which, by monotonicity of $r$, is non-negative. Also, let $\theta_0 = d_0$ and $\Theta^{\tau}_{n_1} = \{\theta : |\theta - \theta_0| \leq K_{\tau} n_1^{-1/3}\}$ where $K_{\tau}$ is chosen such that $P\left(\hat{\theta}_{n_1} \in \Theta^{\tau}_{n_1}\right) > 1 - \tau$ for $\tau > 0$. As $\gamma < 1/3$, 0 is contained in all the intervals $\mathcal{D}_{\theta}$, $\theta \in \Theta^{\tau}_{n_1}$ (equivalently, $d_0 \in [\theta - K n_1^{-\gamma}, \theta + K n_1^{-\gamma}]$), eventually. Note that $M_{n_2}(0, \theta) = 0$. Hence, 0 is a minimizer of $M_{n_2}(\cdot, \theta)$ over $\mathcal{D}_{\theta}$ for each $\theta \in \Theta^{\tau}_{n}$. The process $M_{n_2}$ is a population equivalent of $\mathbb{M}_{n_2}$ and hence, $\hat{u}$ estimates 0. We have the following result for the rate of convergence of $\hat{u}$.

**Theorem 6.10.** *Assume that $r$ is continuously differentiable in a neighborhood of $d_0$ with $r'(d_0) \neq 0$. Then, for $\alpha = (1 - 2\gamma)/3$, $n_2^{\alpha} \hat{u} = O_p(1)$.*

The proof is given in Section E.6 of Appendix E. Next, we derive the limiting distribution of $\hat{d}_2$ by studying the limiting behavior of $\hat{w} = n_2^{\alpha} \hat{u}$. Let $f_{n_2,w,\theta} = {n_2}^{1/6-4\gamma/3} g_{n_2, w n_2^{-\alpha}, \theta}$, $\zeta_{n_2}(w, \theta) = \sqrt{n_2} P f_{n_2,w,\theta}$ and

$$Z_{n_2}(w, \theta) = \mathbb{G}_{n_2} f_{n_2,w,\theta} + \zeta_{n_2}(w, \theta).$$

Then, $n_2^{\alpha} \hat{u} = \hat{w} = \text{argmin}_{w : n_2^{-\alpha} w \in \mathcal{D}_{\hat{\theta}_{n_1}}} Z_{n_2}(w, \hat{\theta}_{n_1})$. We have the following result for the weak convergence of $Z_{n_2}$.

**Theorem 6.11.** *Let $B$ be a standard Brownian motion on $\mathbb{R}$ and*

$$Z(w) = \sigma \sqrt{\frac{p^{\gamma}}{2K(1-p)^{\gamma}}} B(w) + \left(\frac{p}{1-p}\right)^{\gamma} \frac{r'(d_0)}{4K} w^2.$$

*The processes $Z_{n_2}(w, \hat{\theta}_{n_1})$ are asymptotically tight and converge weakly to $Z$. Further,*

$$n^{(1+\gamma)/3}(\hat{d}_2 - d_0) \xrightarrow{d} \left(\frac{8\sigma^2 K}{(r'(d_0))^2 p^{\gamma}(1-p)}\right)^{1/3} \underset{w}{\text{argmin}}\{B(w) + w^2\}. \tag{6.26}$$

The proof is given in Section E.7 of Appendix E. As was the case with the change-point problem, extensions of the above result to non-uniform random designs are possible as well. Also, the proportion $p$ can be optimally chosen to minimize the limiting variance of the second stage estimate. More details on this and related implementation issues can be found in Tang et al. (2013, Section 2.4).

### 6.3.1 Application to a classification problem

In this section, we study a non-parametric classification problem where we show that a multi-stage procedure yields a better classifier in the sense of approaching the misclassification rate of the Bayes classifier.

Consider a model $Y \sim Ber(r(X))$, where $r(x) = P(Y = 1 \mid X = x)$ is a function on $[0, 1]$ and the experimenter has freedom to choose the design distribution (distribution of $X$). Interest centers on using the training data $\{Y_i, X_i\}_{i=1}^n$, obtained from an experimental design setting, to develop a classifier that predicts $Y$ at a given realization $X = x$. A classifier $f$ is a simply a function from $[0, 1]$ to $\{0, 1\}$ which provides a decision rule; assign $x$ to the class $f(x)$. The misclassification rate or the risk of a classifier $f$ is given by

$$\mathcal{R}(f) = P[Y \neq f(X)] .$$

As $\mathcal{R}(f) = E[P[Y \neq f(X) \mid X]]$ which equals

$$E[1[f(X) = 0] r(X) + 1[f(X) = 1] (1 - r(X))] ,$$

it is readily shown that $\mathcal{R}(f)$ is at its minimum for the Bayes classifier $f^*(x) = 1[r(x) \geq 1/2]$. Note that the Bayes classifier cannot be computed simply from the data as $r(\cdot)$ is unknown. It is typical to evaluate the performance of a classifier $f$ by comparing its (asymptotic) risk to that of the Bayes classifier which is the best performing decision rule in terms of $\mathcal{R}(\cdot)$.

We study the above model under the shape-constraint that $r(\cdot)$ is monotone. In this setting, $r^{-1}(1/2)$ can be estimated in an efficient manner through the multistage procedure spelled out in Section 6.3. Let $\hat{d}_2 = \hat{r}_{n_2}^{-1}(1/2)$ denote the second stage estimate. In contrast with Section 6.3, we now have a binary regression model with the underlying regression function being monotone. It is noted here that the asymptotic results for $\hat{d}_2$ in this model parallel those for a heteroscedastic isotonic regression model (note that $\mathrm{Var}(Y \mid X) = r(x)(1 - r(x))$) and can be established in analogous manner. For example, it can be shown that

$$n^{(1+\gamma)/3}(\hat{d}_2 - d_0) \xrightarrow{d} \left( \frac{8Kr(d_0)(1 - r(d_0))}{(r'(d_0))^2 p^\gamma (1 - p)} \right)^{1/3} \underset{w}{\operatorname{argmin}}\{B(w) + w^2\}, \qquad (6.27)$$

where $d_0 = r^{-1}(1/2)$. Here, the variance $\sigma^2$ in Theorem 6.11 gets replaced by $\mathrm{Var}(Y \mid X = d_0) = r(d_0)(1 - r(d_0))$.

Now, an efficient classifier can be constructed as

$$\hat{f}(x) = 1\left[\hat{r}_{n_2}(x) \geq 1/2\right] = 1\left[x \geq \hat{d}_2\right].$$

We study the limiting risk of this classifier with that for the Bayes rule $f^*$. To fix ideas, we define $\mathcal{R}(\cdot)$ with respect to the first stage design distribution which we take to be uniform on $[0, 1]$ for simplicity. Note that the Bayes classifier is invariant of the design distribution and hence, a valid classifier to compare with.

We have the following result on the misclassification rate of $\hat{f}$. Here, $\mathcal{R}(\hat{f})$ is interpreted as $\mathcal{R}(f)$ computed at $f = \hat{f}$.

**Theorem 6.12.** *Assume that $r$ is continuously differentiable in a neighborhood of $d_0$ with $r'(d_0) \neq 0$. Then,*

$$n^{2(1+\gamma)/3}(\mathcal{R}(\hat{f}) - \mathcal{R}(f^*)) \xrightarrow{d} \left( \frac{8Kr(d_0)(1 - r(d_0))}{\sqrt{r'(d_0)} p^\gamma (1 - p)} \right)^{2/3} \left[ \underset{w}{\operatorname{argmin}}\{B(w) + w^2\} \right]^2.$$

This is a significant improvement over the corresponding single stage procedure, the procedure that is equivalent to working with the first-stage classifier $\tilde{f} = 1[x \geq \hat{d}_1]$

($\hat{d}_1$ is the first stage estimate) whose risk approaches the Bayes risk at the rate $n_1^{2/3}$.

**Theorem 6.13.** *Assume that $r$ is continuously differentiable in a neighborhood of $d_0 = r^{-1}(1/2)$. Then,*

$$n_1^{2/3}(\mathcal{R}(\tilde{f}) - \mathcal{R}(f^*)) \xrightarrow{d} \left( \frac{4r(d_0)(1-r(d_0))}{\sqrt{r'(d_0))}} \right)^{2/3} \left[ \underset{w}{\operatorname{argmin}}\{B(w) + w^2\} \right]^2 .$$

We prove Theorem 6.12 below. The proof of Theorem 6.13 follows along the same lines starting from the limit distribution of $\hat{d}_1$.

*Proof of Theorem 6.12.* Note that for $f(x) = 1\,[x \geq a]$

$$\mathcal{R}(f) = \int_0^a r(x)dx + \int_a^1 (1-r(x))dx = \int_0^1 (1-r(x))dx + \int_0^a (2r(x)-1)dx.$$

For notational ease, we use $\int_c^d$ to denote $-\int_d^c$ whenever $c > d$. Then, by a change of variable,

$$\begin{aligned} n^{2(1+\gamma)/3}(\mathcal{R}(\hat{f}) - \mathcal{R}(f^*)) &= n^{(1+\gamma)/3} \int_{d_0}^{\hat{d}_2} 2(r(x) - 1/2)dx \\ &= n^{(1+\gamma)/3} \int_0^{\left(n^{(1+\gamma)/3}(\hat{d}_2 - d_0)\right)} 2(r(d_0 + hn^{-(1+\gamma)/3}) - r(d_0))dh. \end{aligned}$$

By Skorokhod's representation theorem, a version of $n^{(1+\gamma)/3}(\hat{d}_2 - d_0)$, say $\xi_n(\omega)$, converges almost surely to a tight random variable $\xi(\omega)$ which has the same distribution as the random variable on right side of (6.27). As $r$ is continuously differentiable in a neighborhood of $d_0 = r^{-1}(1/2)$, there exists $\delta_0 > 0$, such that $|r'(x)| < 2r'(d_0)$, whenever $|x - d_0| < \delta_0$. Hence, for a $\tau > 0$ and a fixed $\omega$, there exist $N_{\omega,\tau,\delta_0} \in \mathbb{N}$, such that $|\xi_n(\omega) - \xi(\omega)| < \tau$ and $(|\xi(\omega)| + \tau)n^{-(1+\gamma)/3} < \delta_0$ whenever $n > N_{\omega,\tau,\delta_0}$. Hence, for $n > N_{\omega,\tau,\delta_0}$,

$$\begin{aligned}
& n^{(1+\gamma)/3} \int_0^{\xi_n(\omega)} 2(r(d_0 + hn^{-(1+\gamma)/3}) - r(d_0))dh \\
= \; & n^{(1+\gamma)/3} \int_0^{\xi_n(\omega)} 2(r(d_0 + hn^{-(1+\gamma)/3}) - r(d_0))1\left[|h| \leq |\xi(\omega)| + \tau\right] dh \\
= \; & \int_0^{\xi_n(\omega)} (2r'(d_h^\star)h)1\left[|h| \leq |\xi(\omega)| + \tau\right] dh,
\end{aligned}$$

where $d_h^\star$ is an intermediate point between $d_0$ and $d_0 + hn^{-(1+\gamma)/3}$. Note that $r'(d_h^\star)$ converges (pointwise in $h$) to $r'(d_0)$. As the integrand is bounded by $4r'(d_0)h1\left[|h| \leq |\xi(\omega)| + \tau\right]$ which is integrable, by the dominated convergence theorem, the above display then converges to $r'(d_0)\xi^2(\omega)$. Consequently,

$$P\left(n^{(1+\gamma)/3} \int_0^{\xi_n} 2(r(d_0 + hn^{-(1+\gamma)/3}) - r(d_0))dh \not\to r'(d_0)\xi^2\right) \leq P\left(\xi_n \not\to \xi\right) = 0.$$

Thus, we get the result. $\square$

## 6.4 A mode estimation problem

Consider a model of the form $Y = m(X) + \epsilon$ in an experimental design setting where $m(x) = \tilde{m}(|x - d_0|)$ with $\tilde{m} : [0, \infty) \mapsto \mathbb{R}$ being a monotone decreasing function. Consequently, the regression function $m$ is unimodal and symmetric around $d_0$. Interest centers on estimating the point of maximum $d_0 \in (0, 1)$ which can be thought of as a target or a source emanating signal isotropically in all directions. We assume that $\tilde{m}'(0) < 0$. We propose the following two-stage approach.

1. At stage one, sample $n_1 = pn$ ($p \in (0, 1)$) covariate values uniformly from

$[0,1]$ and, from the obtained data, $(Y_i^{(1)}, X_i^{(1)})_{i=1}^{n_1}$, estimate $d_0$ by $\hat{d}_1 = \text{argmax}_{d\in(b,1-b)} \mathbb{M}_{n_1}(d)$, where

$$\mathbb{M}_{n_1}(d) = \mathbb{P}_{n_1} Y^{(1)} 1\left[|X^{(1)} - d| \leq b\right], \tag{6.28}$$

where the bin-width $b > 0$ is sufficiently small so that $[d_0 - b, d_0 + b] \subset (0, 1)$. Note that the estimate is easy to compute as the search for the maximum of $\mathbb{M}_{n_1}$ can be restricted to points $d$ such that either $d - b$ or $d + b$ is a design point.

2. For $K > b > 0$ and $\gamma > 0$, sample the remaining $n_2 = (1-p)n$ covariate-response pairs $\{Y_i^{(2)}, X_i^{(2)}\}$, where

$$Y_i^{(2)} = m(X_i^{(2)}) + \epsilon_i^{(2)}, \quad X_i^2 \sim \text{Uniform}[\hat{d}_1 - K{n_1}^{-\gamma}, \hat{d}_1 + K{n_1}^{-\gamma}].$$

Obtain an updated estimate of $d_0$ by

$$\hat{d}_2 = \underset{d\in\mathcal{D}_{\hat{\theta}_{n_1}}}{\text{argmax}}\, \mathbb{M}_{n_2}(d), \text{ where}$$

$$\mathbb{M}_{n_2}(d) = \mathbb{P}_{n_2} Y^{(2)} 1\left[|X^{(2)} - d| \leq b n_1^{-\gamma}\right], \tag{6.29}$$

$\hat{\theta}_{n_1} = \hat{d}_1$ and $\mathcal{D}_{\hat{\theta}_{n_1}} = [\hat{\theta}_{n_1} - (K-b)n_1^{-\gamma}, \hat{\theta}_{n_1} + (K-b)n_1^{-\gamma}]$. Here, $\gamma$ is chosen such that $P\left(d_0 \in [\hat{d}_1 - (K-b){n_1}^{-\gamma}, \hat{d}_1 + (K-b){n_1}^{-\gamma}]\right)$ converges to 1. It will be shown that ${n_1}^{1/3}(\hat{d}_1 - d_0) = O_p(1)$. Hence, any choice of $\gamma < 1/3$ suffices.

The single stage approach is adapted from the shorth procedure (see, for example, Kim and Pollard (1990, Section 6)) originally developed to find the mode of a symmetric density. The limiting behavior of the first stage estimate is derived next.

**Theorem 6.14.** *We have* ${n_1}^{1/3}(\hat{d}_1 - d_0) = O_p(1)$ *and*

$$ {n_1}^{1/3}(\hat{d}_1 - d_0) \overset{d}{\Rightarrow} \left(\frac{a}{c}\right)^{2/3} \operatorname{argmax}\left\{B(h) - h^2\right\} \tag{6.30} $$

*where* $a = \sqrt{2(m^2(d_0 + b) + \sigma^2)}$ *and* $c = -m'(d_0 + b) > 0$.

**Remark 6.15.** *The symmetry of the function* $m$ *around* $d_0$ *is necessary. If* $m$ *were not symmetric, our procedure (at the first stage), which reports the center of the bin (with width* $2b$*) having the maximum average height as the estimate of* $d_0$*, need not be consistent. For example, when* $m(x) = \exp(-a_1|x - d_0|)$ *for* $x \leq d_0$*, and* $m(x) = \exp(-a_2|x - d_0|)$ *for* $x > d_0$*,* $(a_1 \neq a_2)$*, elementary calculations show that the expected criterion function,* $E\left[\mathbb{M}_{n_1}(d)\right]$ *is minimized at* $d^* = d_0 + (a_1 - a_2)b/(a_1 + a_2) \neq d_0$ *and that* $\hat{d}_1$ *is a consistent estimate of* $d^*$.

The proof follows from application of standard empirical process results and is outlined in Section E.8. For the second stage, we get the following result.

**Theorem 6.16.** *We have* ${n_2}^{(1+\gamma)/3}(\hat{d}_2 - d_0) = O_p(1)$ *and*

$$ n^{(1+\gamma)/3}(\hat{d}_2 - d_0) \xrightarrow{d} \left(\frac{4K(m^2(d_0) + \sigma^2)}{(m'(d_0+))^2 p^{\gamma}(1-p)}\right)^{1/3} \operatorname{argmax}\left\{B(h) - h^2\right\} \tag{6.31} $$

**Remark 6.17.** *It is critical here to work with a uniform design for this problem. The uniform design at each stage ensures that the population criterion function is maximized at the true parameter* $d_0$*. In fact, if a non-flat random design is used at the second stage with design distribution symmetric at* $\hat{d}_1$*, it can be shown that* $\hat{d}_2$ *can not converge at a rate faster than* $n^{1/3}$ *as it effectively ends up estimating an intermediate point between* $d_0$ *and* $\hat{d}_1$*. Further, if a non-flat design is used at the first stage, it can be shown that* $\hat{d}_1$ *need not be consistent for* $d_0$*.*

**Remark 6.18.** *Root finding algorithms (Robbins and Monro, 1951) and their extensions (Kiefer and Wolfowitz, 1952) provide a classical approach for locating the* *maximum of the a regression function in an experimental design setting. However, due to the non-smooth nature of our problem ($m$ not being differentiable at $d_0$), $d_0$ is no longer the solution to the equation $m'(d) = 0$, and hence, these algorithms do not apply to our setting.*

The proof is given in Section E.9 of Appendix E. As was the case with the change-point and the inverse isotonic regression problem, an optimal choice for the proportion $p$ exists that minimizes the limiting variance of the second stage estimate. As before, $K$ and $\gamma$ are chosen in practice such that

$$K{n_1}^{-\gamma} \approx \frac{C_{\tau/2}}{n_1^{1/3}},$$

with $C_{\tau/2}$ being the $(1-\tau/2)$th quantile of the limiting distribution of $n_1^{1/3}(\hat{d}_1 - d_0)$. The variance of $(\hat{d}_2 - d_0)$ would be (approximately) at its minimum when

$$\frac{1}{n^{(1+\gamma)/3}}\left(\frac{4K(m^2(d_0)+\sigma^2)}{(m'(d_0+))^2 p^{\gamma}(1-p)}\right)^{1/3} \approx \frac{1}{n^{4/9}}\left(\frac{4C_{\tau/2}(m^2(d_0)+\sigma^2)}{(m'(d_0+))^2 p^{1/3}(1-p)}\right)^{1/3}$$

is at its minimum. Equivalently, $p^{1/3}(1-p)$ needs to be at its maximum. This yields the optimal choice of $p$ to be $p_{opt} \approx 0.25$.

## 6.5 Conclusions

*Negative examples and possible solutions.* In this chapter, we considered examples where multistage procedures accentuated the efficiency of the M-estimates by accelerating the rate of convergence. However, this is not a universal phenomenon. In most regular parametric problems, where the estimates exhibit a $\sqrt{n}$-rate of convergence, acceleration to a faster rate is not possible. Also, in the mode estimation problem considered in this chapter, it can be shown that if the regression function is smooth at $d_0$, i.e., $m'(d_0) = 0$, the second stage estimate converges at a slower rate than the first stage estimate. This is due to the fact that the function appears almost flat in the (second stage) zoomed-in neighborhood and our criterion that simply relies on finding the bin with maximum average height is not able to capture the local quadratic nature of the function in this shrinking neighborhood. In such a situation quite a few extensions are possible. Working with a symmetric (non-flat) design centered at the first stage estimate, which affects the population criterion function favorably in this setting (in contrast with Remark 6.17), an $n^{1/3}$-rate of convergence can be maintained for the second stage estimate. Alternatively, one can fit a quadratic curve (which is the local nature of the regression function $m$, provided $m''(d_0) \neq 0$) at the second stage which is expected to accelerate the rate of convergence. Some work in this direction can be found in Hotelling (1941). As mentioned earlier, similar phenomena were observed in Tang et al. (2011), where they constructed a $\sqrt{n}$-consistent estimate by doing a linear approximation at the second stage of sampling to estimate the inverse of a monotone regression function. We note here that unlike our settings, the shrinking neighborhood chosen in Tang et al. (2011) was not required to contain $d_0$ with high probability. This is due to the fact that their criterion function leverages on the linear approximation of regression function and can extrapolate to estimate $d_0$. Hence, the acceleration in the rate (or the lack of it) turns out to be a feature of the model as well as the method. Further, as mentioned in Remark 6.18, Kiefer-Wolfowitz procedure (Kiefer and Wolfowitz, 1952) can also be used to estimate the location of the maximum of the regression function in this smooth $m$ setting with $m'(d_0) = 0$.

*Pooled data.* In certain models, it is preferred, at least from the perspective of reducing the limiting variance, to pool the data across stages to obtain the final estimates. For example, in change-point models where regression function is linear on either side of the threshold, e.g., $m(x) = (\alpha_0 + \alpha_1 x)1(x \leq d_0) + (\beta_0 + \beta_1 x)1(x > d_0)$, $\alpha_i \neq \beta_i$, $i = 1, 2$, it is recommended to estimate at least the slope parameters using the pooled data. This is due to the fact that slopes are better estimated when the design points are far apart. It is far from clear whether a generic formulation is possible to approach such multistage procedures as the nature of the dependence gets more convoluted.

# Part III

# Random fields

# CHAPTER 7

# A central limit theorem for linear random fields

Random fields have attracted a lot of attention especially in modeling spatially correlated data. They are encountered in several applications from geo-spatial statistics, environmental statistics, human brain mapping, and image processing (e.g., see Cressie (1991); Ivanov and Leonenko (1989)). Limit theorems for random fields have motivated a number of papers. They have been studied under different settings.

In this chapter, we are mainly interested in Central Limit Theorems (CLT) for linear random fields. In the paper by Phillips and Solo (1992), it was demonstrated that the so-called Beveridge-Nelson decomposition (BND) presents a simple method for proving limit theorems for sums of values of linear processes. This has been exploited in several papers to decompose partial sums of linear random fields into a partial sum of independent components and a remainder term which can be dealt more readily. Using this technique, Marinucci and Poghosyan (2001) proved IP for partial sums of linear random fields with independent innovations over rectangles as well as a strong approximation result for the same by a Gaussian random field. This was further generalized for dependent innovations by Ko et al. (2008). Paulauskas (2010) used BND to obtain sufficient conditions for CLT and strong laws for the partial sums of linear random fields as well as their squares summed over sets such as rectangles and squares. Using BND seems to give weak limit results at the price of a simpler proof.

Our approach to the problem does not rely upon the use of BND. We provide sufficient conditions for CLT to hold for linear random fields based upon a criterion that arises naturally starting from the Lindberg-Feller condition. The approach requires innovations to be independent and does not deliver a functional version. However, we allow sums to be taken over sets as general as the disjoint union of rectangles. Also, our result in its simpler form extends a CLT for linear processes (Ibragimov, 1962) to that for linear random fields with no extra conditions.

We formulate the problem in a two-dimensional setting in Section 7.1. The results presented in this chapter can be easily extended to $d$-dimensional, linear random fields, $d \geq 1$. Notational ease restricts us to illustrate the techniques for $d = 2$ only. In Section 7.2, we deduce fundamental criteria for CLT to hold when sums are taken over general shapes. Simpler conditions are derived in Section 7.3 which ensure a CLT for linear random fields when sums are taken over finite union of rectangles.

## 7.1 Formulation

Consider a two-dimensional, linear random field, say

$$X_{j,k} = \sum_{r\in\mathbb{Z}}\sum_{s\in\mathbb{Z}} a_{r,s}\xi_{j-r,k-s} = \sum_{r\in\mathbb{Z}}\sum_{s\in\mathbb{Z}} a_{j+r,k+s}\xi_{-r,-s},$$

where $a_{r,s},\ r,s \in \mathbb{Z}$, are square summable, $\xi_{r,s},\ r,s \in \mathbb{Z}$, are i.i.d. with mean 0 and unit variance, and $\mathbb{Z}$ denotes the integers. It is convenient to regard the array $\mathbf{a} = (a_{r,s} : r,s \in \mathbb{Z})$ as an element of $\ell^2(\mathbb{Z}^2)$. Let $F$ denote the common distribution function of the $\xi_{r,s}$ and $(\Omega, \mathcal{A}, P)$ the probability space on which they are defined. If $\Gamma$ is a finite subset of $\mathbb{Z}^2$, let

$$S = S(\mathbf{a}, \Gamma) = \sum_{(j,k)\in\Gamma} X_{j,k}$$

and

$$\sigma^2 = \sigma^2(\mathbf{a}, \Gamma) = E(S^2),$$

and suppose that $\sigma^2 > 0$. (Of course, $S$ depends on $\omega \in \Omega$ too, but this dependence is suppressed. As indicated, dependence on $\mathbf{a}$ and $\Gamma$ will only be displayed when needed for clarity.) Then

$$S = \sum_{r \in \mathbb{Z}} \sum_{s \in \mathbb{Z}} b_{r,s} \xi_{-r,-s},$$

where

$$b_{r,s} = b_{r,s}(\mathbf{a}, \Gamma) = \sum_{(j,k) \in \Gamma} a_{j+r,k+s},$$

and

$$\sigma^2 = \sum_{r \in \mathbb{Z}} \sum_{s \in \mathbb{Z}} b_{r,s}^2,$$

assumed to be positive. Let $\Phi$ denote the standard normal distribution and

$$G(z) = G(z; \mathbf{a}, \Gamma, F) = P\left[\frac{S}{\sigma} \leq z\right], \ z \in \mathbb{R}.$$

Sufficient conditions for $G$ to be close to $\Phi$ are developed.

## 7.2 Generalities

Let

$$\rho = \rho(\mathbf{a}, \Gamma) = \max_{r,s \in \mathbb{Z}} \frac{|b_{r,s}|}{\sigma}. \tag{7.1}$$

Interest in $\rho$ stems from the following:

**Proposition 7.1.** *Let $\mathcal{H}$ denote a class of distribution functions for which*

$$\int_{\mathbb{R}} x H\{dx\} = 0, \quad \int_{\mathbb{R}} x^2 H\{dx\} = 1, \text{for all } H \in \mathcal{H}, \tag{7.2}$$

$$\lim_{c\to\infty} \sup_{H\in\mathcal{H}} \int\limits_{|x|>c} x^2 H\{dx\} = 0; \tag{7.2}$$

*Then* $\forall\ \epsilon > 0$, $\exists\ \delta = \delta_{\epsilon,\mathcal{H}}$, *depending only on* $\epsilon$ *and* $\mathcal{H}$ *for which*

$$d(G,\Phi) := \sup_z |G(z) - \Phi(z)| \le \epsilon \tag{7.3}$$

*for all* $F \in \mathcal{H}$ *for all arrays* $\mathbf{a}$ *and finite regions* $\Gamma \subset \mathbb{Z}^2$ *for which* $\rho \le \delta$.

*Proof:* Let $\hat{}$ denote Fourier transform (characteristic function), so that $\hat{F}(t) = \int_{\mathbb{R}} e^{\imath tx} F\{dx\}$, and

$$L(\eta) = \frac{1}{\sigma^2} \sum_{r,s\in\mathbb{Z}} \int\limits_{|b_{r,s}x|>\eta\sigma} |b_{r,s}x|^2 F\{dx\}$$

for $\eta > 0$. Then, for any $\eta > 0$, $|\hat{G}(t) - \hat{\Phi}(t)| \le \eta|t|^3 + t^2 L(\eta) + t^4 \exp(t^2)\rho^2$ for all $t \in (0, \sqrt{2}/\rho)$ from the proof of the Central Limit Theorem for independent summands (Billingsley, 1995, pp. 359-361) and

$$\begin{aligned}\sup_z |G(z) - \Phi(z)| &\le \frac{1}{\pi} \int\limits_{-T}^{T} \left| \frac{\hat{G}(t) - \hat{\Phi}(t)}{t} \right| dt + \frac{24}{\pi\sqrt{2\pi}T} \\ &\le \frac{\eta T^3}{3} + \frac{1}{2} T^2 L(\eta) + \frac{1}{2} e^{T^2}(T^2 - 1)\rho^2 + \frac{24}{T}\end{aligned} \tag{7.4}$$

for any $T \in (0, \sqrt{2}/\rho)$ by the smoothing inequality (Feller, 1971, pp. 510-512). Given $\epsilon > 0$, let $T_\epsilon = 96/\epsilon$ and $\eta_\epsilon = 3\epsilon/(4T_\epsilon^3)$. Then the left side of (7.4) is at most

$$\frac{1}{2} T_\epsilon^2 L(\eta_\epsilon) + \frac{1}{2} e^{T_\epsilon^2}(T_\epsilon^2 - 1)\rho^2 + \frac{1}{2}\epsilon,$$

provided $\rho < \sqrt{2}/T_\epsilon$. Next, let $J(c) = \sup_{H\in\mathcal{H}} \int_{|x|>c} x^2 H\{dx\}$ for $c > 0$, so that $J(c) \to 0$ as $c \to \infty$ by (7.2); and let $J^\#(z) = \inf\{c > 0 : J(c) \le z\}$ for $z > 0$. Then

$$L(\eta) \le \frac{1}{\sigma^2} \sum_{r\in\mathbb{Z}} \sum_{s\in\mathbb{Z}} b_{r,s}^2 \int\limits_{|x|>\eta/\rho} x^2 F\{dx\} \le J\left(\frac{\eta}{\rho}\right),$$

and $\delta = \min\left(\sqrt{2}/T_\epsilon, \sqrt{\epsilon/(2\exp(T_\epsilon^2)(T_\epsilon^2-1))}, \eta_\epsilon/J^\#(T_\epsilon^{-2}\epsilon/2)\right)$ has the desired properties. $\square$

Next, let

$$\|\mathbf{a}\|_p = \left[\sum_{r,s\in\mathbb{Z}} |a_{r,s}|^p\right]^{\frac{1}{p}} \leq \infty$$

for $1 \leq p \leq 2$. Thus, $\|\mathbf{a}\|_2$ is assumed to be finite and $\|\mathbf{a}\|_p$ may be finite for some value of $p < 2$. In terms of $\|\mathbf{a}\|_p$ there is a simple bound on $\rho$,

$$\rho \leq \frac{\|\mathbf{a}\|_p(\#\Gamma)^{\frac{1}{q}}}{\sigma}, \tag{7.5}$$

where $q$ denotes the conjugate, $1/p + 1/q = 1$ and $\#\Gamma$ denotes the cardinality of $\Gamma$. In particular, $\rho \leq \|\mathbf{a}\|_1/\sigma$. This leads to:

**Corollary 7.2.** *Let $\mathcal{H}$ be as in Proposition 1. If $\|\mathbf{a}\|_1 < \infty$, then $\forall\ \epsilon > 0,\ \exists\ \kappa = \kappa_{\epsilon,\mathcal{H}} > 0$ for which (7.3) holds whenever $\sigma \geq \kappa\|\mathbf{a}\|_1$ and $F \in \mathcal{H}$.*

*Proof:* For $\delta$ as per *Proposition 1*, let $\kappa = 1/\delta$. The result is then a consequence of the proposition and the fact that $\rho \leq \|\mathbf{a}\|_1/\sigma$. $\square$

An immediate consequence of the above result is the following.

**Corollary 7.3.** *Let $\|\mathbf{a}\|_1 < \infty$ and $\Gamma_n$ be a sequence of finite subsets of $\mathbb{Z}^2$ such that $\sigma(\mathbf{a}, \Gamma_n) \to \infty$ as $n \to \infty$. Then the distributions of $S(\mathbf{a}, \Gamma_n)/\sigma(\mathbf{a}, \Gamma_n)$ converge to $\Phi$.*

**Remark 7.4.** *Our result does not put any restrictions the shape of the sets $\Gamma_n$ but it requires $\#\Gamma_n \to \infty$ (as a consequence of $\sigma(\mathbf{a}, \Gamma_n) \to \infty$). In Paulauskas (2010), a similar theorem is shown to hold for rectangles, i.e., $\Gamma_n = \{(j,k) : 1 \leq j \leq M(n), 1 \leq k \leq N(n)\}$, with $\min(M(n), N(n)) \to \infty$. But it was not resolved whether a CLT would hold with the weaker condition $\#\Gamma_n \to \infty$. In particular, our result provides a CLT when $\Gamma_n$s are effectively "one-dimensional", e.g., $M(n) \equiv 1$ and $N(n) \to \infty$.*

## 7.3 Union of Rectangles

We specialize our results to $\Gamma$ being a union of finitely many discrete rectangles. To bound $\rho$, suppose that the maximum in (7.1) occurs when $r = r_0$ and $s = s_0$, say

$$|b_{r_0,s_0}| = \max_{r,s} |b_{r,s}|,$$

and let $\Delta b_{u,v} = b_{u,v} - b_{u,v-1} - b_{u-1,v} + b_{u-1,v-1}$ for $(u, v) \in \mathbb{Z}^2$. Then

$$b_{r_0+r,s_0+s} - b_{r_0,s_0+s} - b_{r_0+r,s_0} + b_{r_0,s_0} = \sum_{u=r_0+1}^{r_0+r} \sum_{v=s_0+1}^{s_0+s} \Delta b_{u,v} \tag{7.6}$$

for $r, s \geq 1$. Let

$$Q_{m,n} = \sum_{r=1}^{m} \sum_{s=1}^{n} \sum_{u=r_0+1}^{r_0+r} \sum_{v=s_0+1}^{s_0+s} |\Delta b_{u,v}| = \sum_{r=r_0+1}^{r_0+m} \sum_{s=s_0+1}^{s_0+n} (r - r_0)(s - s_0)|\Delta b_{r,s}| \tag{7.7}$$

for $m, n \geq 1$. Since $b_{r_0,s_0} = -b_{r_0+r,s_0+s} + b_{r_0,s_0+s} + b_{r_0+r,s_0} + \sum_{u=r_0+1}^{r_0+r} \sum_{v=s_0+1}^{s_0+s} \Delta b_{u,v}$, we have $|b_{r_0,s_0}| \leq |b_{r_0,s_0+s}| + |b_{r_0+r,s_0}| + |b_{r_0,s_0}| + \sum_{u=r_0+1}^{r_0+r} \sum_{v=s_0+1}^{s_0+s} |\Delta b_{u,v}|$ for all $r, s \geq 1$ and, therefore,

$$mn|b_{r_0,s_0}| \leq \sum_{r=1}^{m} \sum_{s=1}^{n} (|b_{r_0+r,s_0+s}| + |b_{r_0,s_0+s}| + |b_{r_0+r,s_0}|) + Q_{m,n}$$

for all $m, n \geq 1$. Here

$$\sum_{r=1}^{m} \sum_{s=1}^{n} |b_{r_0+r,s_0+s}| \leq \sqrt{mn} \sqrt{\sum_{r=1}^{m} \sum_{s=1}^{n} b^2_{r_0+r,s_0+s}} \leq \sqrt{mn}\sigma,$$

and similarly, $\sum_{r=1}^{m} \sum_{s=1}^{n} |b_{r_0,s_0+s}| \leq m\sqrt{n}\sigma$ and $\sum_{r=1}^{m} \sum_{s=1}^{n} |b_{r_0+r,s_0}| \leq \sqrt{m}n\sigma$. So, $mn|b_{r_0,s_0}| \leq \sqrt{mn}\sigma + m\sqrt{n}\sigma + \sqrt{m}n\sigma + Q_{m,n}$. That is,

$$\rho = \frac{|b_{r_0,s_0}|}{\sigma} \leq (\frac{2}{\sqrt{m}} + \frac{2}{\sqrt{n}}) + \frac{Q_{m,n}}{mn\sigma}, \tag{7.8}$$

for any $m, n \geq 1$. The first two terms can be made small by taking $m$ and $n$ large. Thus, the issue is $Q_{m,n}$. Suppose now that $\Gamma$ can be written as the union of $\ell$ non-empty pairwise mutually exclusive rectangles, i.e.,

$$\Gamma = \bigcup_{i=1}^{\ell} \{(j,k) : \underline{M}_i \leq j \leq \overline{M}_i,\ \underline{N}_i \leq k \leq \overline{N}_i\}. \tag{7.9}$$

**Proposition 7.5.** *If $\Gamma$ is of the form (7.9), then*

$$\rho \leq 20 \left( \frac{\sqrt{\ell}\|\mathbf{a}\|_2}{\sigma} \right)^{\frac{1}{5}} + \frac{8\sqrt{\ell}\|\mathbf{a}\|_2}{\sigma}. \tag{7.10}$$

*Proof:* In this case $b_{r,s} = \sum_{i=1}^{\ell} b_{r,s}^{(i)}$, where $b_{r,s}^{(i)} = \sum_{j=\underline{M}_i}^{\overline{M}_i} \sum_{k=\underline{N}_i}^{\overline{N}_i} a_{j+r,k+s}$, $\Delta b_{r,s} = \sum_{i=1}^{\ell} \Delta b_{r,s}^{(i)}$, and $\Delta b_{r,s}^{(i)} = a_{r+\overline{M}_i,s+\overline{N}_i} - a_{r+\underline{M}_i,s+\overline{N}_i} - a_{r+\overline{M}_i,s+\underline{N}_i} + a_{r+\underline{M}_i,s+\underline{N}_i}$. So,

$$\begin{aligned} Q_{m,n} &\leq mn \sum_{r=1}^{m} \sum_{s=1}^{n} \sum_{i=1}^{\ell} (|a_{r+\overline{M}_i,s+\overline{N}_i} - a_{r+\underline{M}_i,s+\overline{N}_i} - a_{r+\overline{M}_i,s+\underline{N}_i} + a_{r+\underline{M}_i,s+\underline{N}_i}|) \\ &\leq 4(mn)^2\sqrt{\ell}\|\mathbf{a}\|_2, \end{aligned}$$

by Schwartz' Inequality, and

$$\rho \leq (\frac{2}{\sqrt{m}} + \frac{2}{\sqrt{n}}) + \frac{4mn\sqrt{\ell}\|\mathbf{a}\|_2}{\sigma},$$

for any $m, n \geq 1$. Letting $m = n = \lceil (\sigma/\sqrt{\ell}\|\mathbf{a}\|_2)^{\frac{2}{5}} \rceil$, the least integer that exceeds $(\sigma/\sqrt{\ell}\|\mathbf{a}\|_2)^{\frac{2}{5}}$, leads to (7.10). $\square$

When specialized to (intersections of) rectangles (with $\mathbb{Z}^2$), we have the following result.

**Corollary 7.6.** *Let $\mathcal{H}$ be as in Proposition 7.1 and let $\mathcal{R}_\kappa$ be the collection of pairs $(\mathbf{a}, \Gamma)$ for which $\|\mathbf{a}\|_2 > 0$, $\Gamma$ is a finite rectangle, and $\sigma(\mathbf{a}, \Gamma) \geq \kappa\|\mathbf{a}\|_2$. Then, as $\kappa \to \infty$, the distributions of $S/\sigma$ converge to $\Phi$ uniformly with respect to $(\mathbf{a}, \Gamma) \in \mathcal{R}_\kappa$*

*and* $F \in \mathcal{H}$.

This provides a complete analogue of Ibragimov's theorem (Ibragimov, 1962), with a lot of uniformity. Next, we give a convenient formulation which, like *Corollary 7.3*, provides a CLT with a weak condition ($\#\Gamma_n \to \infty$) on the growth of rectangles $\Gamma_n$, defined below.

**Corollary 7.7.** *Let* $\Gamma_n = \{(j,k) \ : \ 1 \leq j \leq M(n), 1 \leq k \leq N(n)\}$ *for* $M(n), N(n) \geq 1$. *If* $\sigma(\mathbf{a}, \Gamma_n) \to \infty$ *as* $\#\Gamma_n \to \infty$, *then the distributions of* $S(\mathbf{a}, \Gamma_n)/\sigma(\mathbf{a}, \Gamma_n)$ *converge to* $\Phi$.

# APPENDICES

# APPENDIX A

# Proofs for Chapter 2

We start with establishing an auxiliary result that is used in the subsequent developments.

**Theorem A.1.** *Let $\mathcal{T}$ be an indexing set and $\{\mathbb{M}_n^\tau : \tau \in \mathcal{T}\}_{n=1}^\infty$ a family of real-valued stochastic processes indexed by $h \in \mathcal{H}$. Also, let $\{M^\tau : \tau \in \mathcal{T}\}$ be a family of deterministic functions defined on $\mathcal{H}$, such that each $M^\tau$ is maximized at a unique point $h(\tau) \in \mathcal{H}$. Here $\mathcal{H}$ is a metric space and denote the metric on $\mathcal{H}$ by $d$. Let $\hat{h}_n^\tau$ be a maximizer of $\mathbb{M}_n^\tau$. Assume further that:*
*(a) $\sup_{\tau \in \mathcal{T}} \sup_{h \in \mathcal{H}} |\mathbb{M}_n^\tau(h) - M^\tau(h)| = o_p(1)$, and*
*(b) for every $\eta > 0$, $c(\eta) \equiv \inf_\tau \inf_{h \notin B_\eta(h(\tau))} [M^\tau(h(\tau)) - M^\tau(h)] > 0$, where $B_\eta(h)$ denotes the open ball of radius $\eta$ around $h$.*

*Then, (i) $\sup_\tau d(\hat{h}_n^\tau, h(\tau)) = o_p(1)$. Furthermore, if $\mathcal{T}$ is a metric space and $h(\tau)$ is continuous in $\tau$, then (ii) $\hat{h}_n^{\tau_n} - h(\tau_0) = o_p(1)$, provided $\tau_n$ converges to $\tau_0$. In particular, if the $\mathbb{M}_n^\tau$s themselves are deterministic functions, the conclusions of the theorem hold with the convergence in probability in (i) and (ii) replaced by usual non-stochastic convergence.*

*Proof.* We provide the proof in the case when $\mathcal{H}$ is a sub-interval of the real line, the case that is relevant for our applications. However, there is no essential difference in generalizing the argument to metric spaces - euclidean distances simply need to be replaced by the metric space distance and open intervals by open balls.

Given $\eta > 0$, we need to deal with $P^*\left(\sup_{\tau\in\mathcal{T}} |\hat{h}_n^\tau - h(\tau)| > \eta\right)$, where $P^*$ is the outer probability. The event $A_{n,\eta} \equiv \{\sup_{\tau\in\mathcal{T}} |\hat{h}_n^\tau - h(\tau)| > \eta\}$ implies that for some $\tau$, $\hat{h}_n^\tau \notin (h(\tau)-\eta, h(\tau)+\eta)$ and therefore $M^\tau(h(\tau)) - M^\tau(\hat{h}_n^\tau) \geq \inf_{h\notin(h(\tau)-\eta,h(\tau)+\eta)}\left[M^\tau(h(\tau)) - M^\tau(h)\right]$. This is equivalent to

$$\begin{aligned}
&M^\tau(h(\tau)) - M^\tau(\hat{h}_n^\tau) + \mathbb{M}_n^\tau(\hat{h}_n^\tau) - \mathbb{M}_n^\tau(h(\tau)) \\
&\geq \inf_{h\notin(h(\tau)-\eta,h(\tau)+\eta)}\left[M^\tau(h(\tau)) - M^\tau(h)\right] + \mathbb{M}_n^\tau(\hat{h}_n^\tau) - \mathbb{M}_n^\tau(h(\tau)).
\end{aligned}$$

Now, $\mathbb{M}_n^\tau(\hat{h}_n^\tau) - \mathbb{M}_n^\tau(h(\tau)) \geq 0$ and the left side of the above inequality is bounded above by

$$2\left\|\mathbb{M}_n^\tau - M^\tau\right\|_{\mathcal{H}} \equiv 2\sup_{h\in\mathcal{H}}\left|\mathbb{M}_n^\tau(h) - M^\tau(h)\right|,$$

implying that $2\|\mathbb{M}_n^\tau - M^\tau\|_{\mathcal{H}} \geq \inf_{h\notin(h(\tau)-\eta,h(\tau)+\eta)}\left[M^\tau(h(\tau)) - M^\tau(h)\right]$ which, in turn, implies that $2\sup_{\tau\in\mathcal{T}}\|\mathbb{M}_n^\tau - M^\tau\|_{\mathcal{H}} \geq \inf_{\tau\in\mathcal{T}}\inf_{h\notin(h(\tau)-\eta,h(\tau)+\eta)}\left[M^\tau(h(\tau)) - M^\tau(h)\right] \equiv c(\eta)$ by definition. Hence $A_{n,\eta} \subset \{\sup_{\tau\in\mathcal{T}}\|\mathbb{M}_n^\tau - M^\tau\|_{\mathcal{H}} \geq c(\eta)/2\}$. By assumptions (a) and (b), $P^*\left(\sup_{\tau\in\mathcal{T}}\|\mathbb{M}_n^\tau - M^\tau\|_{\mathcal{H}} \geq c(\eta)/2\right)$ goes to 0 and therefore so does $P^*(A_{n,\eta})$. □

**Remark A.2.** *We will call the sequence of steps involved in deducing the inclusion:*

$$\left\{\sup_{\tau\in\mathcal{T}}|\hat{h}_n^\tau - h(\tau)| > \eta\right\} \subset \left\{\sup_{\tau\in\mathcal{T}}\|\mathbb{M}_n^\tau - M^\tau\|_{\mathcal{H}} \geq c(\eta)/2\right\},$$

*as* generic steps. *Very similar steps will be required again in the proofs of the theorems to follow. We will not elaborate those arguments, but refer back to the* generic steps *in such cases.*

## A.1 Proof of Theorem 2.1

To exhibit the dependence on the baseline value $\tau_0$ (or its estimate), we use notations of the form $\mathbb{M}_n(d, \tau_0)$ and $\hat{d}_{m,n}(\tau_0)$. For convenience, let $T^{(m)}(X_i) = \sqrt{m}(\bar{Y}_{i\cdot} - \tau_0)$ and $Z_{im}(\tau_0) = \tilde{p}_{m,n}(X_i, \tau_0) = 1 - \Phi(T^{(m)}(X_i))$. As $m$ changes, the distribution of $Z_{im}(\tau_0)$ changes, and so we effectively have a triangular array $\{(X_i, Z_{im}(\tau_0))\}_{i=1}^n \sim P_m$, say. Using empirical process notation, $\mathbb{M}_{m,n}(d, \tau_0) \equiv \mathbb{P}_{n,m}\{Z_{1m}(\tau_0) - 1/4\}1(X_1 \leq d)$, where $\mathbb{P}_{n,m}$ denotes the empirical measure of the data. Firstly, we find the limiting process for $\mathbb{M}_{m,n}(d, \tau_0)$. Define $M_m(d) \equiv P_m\{Z_{1m}(\tau_0) - 1/4\}1(X_1 \leq d)$ where $M_m(d)$ can be simplified as

$$M_m(d) = \int_0^d \{\nu_m(x) - 1/4\} f(x) dx, \tag{A.1}$$

where $\nu_m(x) = E[Z_{im}(\tau_0) \mid X_i = x]$. Observe that for $X_i = x$, as $m \to \infty$, $T^{(m)}(x)$ converges in distribution to $N(0, \sigma^2(x))$ for $x \leq d_0$ and $T^{(m)}(x) = \sqrt{m}\{\bar{Y}_{i\cdot} - \mu(x)\} + \sqrt{m}(\mu(x) - \tau_0) \to \infty$, in probability, for $x > d_0$. Thus, $\nu_m(x) \to \nu(x)$ for all $x \in [0, 1]$, where $\nu(x) = (1/2)1(x \leq d_0)$. Let $M(d)$ be the same expression for $M_m(d)$ in (A.1) with $\nu_m(x)$ replaced by $\nu(x)$, e.g., $M(d) = \int_0^d \{\nu(x) - 1/4\} f(x) dx$. Observe that for $c = (1/4) \int_0^{d_0} f(x) dx$, $M(d) \leq c$ for all $d$, and $M(d_0) = c$. Also, it is easy to see that $d_0$ is the unique maximizer of $M(d)$. Now, the difference $|M_m(d) - M(d)|$, can be bounded by $\int_0^1 |\nu_m(x) - \nu(x)| f(x) dx$ which goes to 0 by the dominated convergence theorem. As the bound does not depend on $d$, we get $\|M_m - M\|_\infty \to 0$, where $\|\cdot\|_\infty$ denotes the supremum. By Theorem A.1, $d_m = \arg\max_{d \in [0,1]} M_m(d) \to \arg\max_{d \in [0,1]} M(d) = d_0$ as $m \to \infty$. It would now suffice to show that $(\hat{d}_{m,n}(\hat{\tau}) - d_m)$ is $o_p(1)$.

Fix $\epsilon > 0$ and consider the event $\{|\hat{d}_{m,n}(\hat{\tau}) - d_m| > \epsilon\}$. Since $d_m$ maximizes $M_m$ and $\hat{d}_{m,n}(\hat{\tau})$ maximizes $\mathbb{M}_{m,n}(\cdot, \hat{\tau})$, by arguments analogous to the *generic steps* in the proof of Theorem A.1, we have:

$$|\hat{d}_{m,n}(\hat{\tau}) - d_m| > \epsilon \Rightarrow \|\mathbb{M}_{m,n}(\cdot, \hat{\tau}) - M_m(\cdot)\|_\infty \geq \eta_m(\epsilon)/2\,,$$

where $\eta_m(\epsilon) = \inf_{d \in (d_m - \epsilon, d_m + \epsilon)^c}\{M_m(d_m) - M_m(d)\}$.

We claim that there exists $\eta > 0$ and an integer $M_0$ such that $\eta_m(\epsilon) > \eta > 0$ for all $m \geq M_0$. To see this, let us bound $M_m(d_m) - M_m(d)$ below by $-2\|M_m - M\|_\infty + M(d_m) - M(d)$. As $\|M_m - M\|_\infty \to 0$ as $m \to \infty$, it is enough to show that there exists $\eta > 0$ such that for all sufficiently large $m$, $\inf_{d \in (d_m - \epsilon, d_m + \epsilon)^c}\{M(d_m) - M(d)\} > \eta$. We split $M(d_m) - M(d)$ into two parts as $\{M(d_0) - M(d)\} + \{M(d_m) - M(d_0)\}$. Notice that by the continuity of $M(\cdot)$, the second term goes to 0. To handle the first term, notice that $M(d)$ is a continuous function with a unique maximum at $d_0$. There exists $M_0 \in \mathbb{N}$ such that for all $m > M_0$, we have $(d_0 - \epsilon/2, d_0 + \epsilon/2) \subset (d_m - \epsilon, d_m + \epsilon)$ as $d_m \to d_0$. So, for $m > M_0$, $\inf_{d \in (d_m - \epsilon, d_m + \epsilon)^c}\{M(d_0) - M(d)\} \geq \inf_{d \in (d_0 - \epsilon/2, d_0 + \epsilon/2)^c}\{M(d_0) - M(d)\}$. As $M(d_0) - M(d)$ is continuous, this infimum is attained in the compact set $[0,1] \cap (d_0 - \epsilon/2, d_0 + \epsilon/2)^c$ and is strictly positive. Thus, a positive choice for $\eta$, as claimed, is available.

The claim yields,

$$\begin{aligned} &P_m(|\hat{d}_{m,n}(\hat{\tau}) - d_m| > \epsilon) \qquad\qquad (A.2)\\ &\quad \leq \; P_m(\|\mathbb{M}_{m,n}(\cdot, \hat{\tau}) - \mathbb{M}_{m,n}(\cdot, \tau_0)\|_\infty > \eta/4) + P_m(\sup_{l \geq n}\|\mathbb{M}_{m,l}(\cdot, \tau_0) - M_m\|_\infty > \eta/4). \end{aligned}$$

For the first term, notice that, $\|\mathbb{M}_{m,n}(\cdot, \hat{\tau}) - \mathbb{M}_{m,n}(\cdot, \tau_0)\|_\infty \leq \max_{i \leq n}|Z_{im}(\hat{\tau}) - Z_{im}(\tau_0)|$. This is bounded above by

$$\sup_{u \in \mathbb{R}}\left|\Phi\left(u\right) - \Phi\left(u + \sqrt{m}(\hat{\tau} - \tau_0)\right)\right|.$$

As $\sup_{u \in \mathbb{R}}|\Phi\left(u\right) - \Phi\left(u + a\right)| = 2\Phi\left(|a|/2\right) - 1$, for $a \in \mathbb{R}$, $\|\mathbb{M}_{m,n}(\cdot, \hat{\tau}) - \mathbb{M}_{m,n}(\cdot, \tau_0)\|_\infty$

is bounded by $\{2\Phi(\sqrt{m}|\hat{\tau} - \tau_0|/2) - 1\}$, which goes in probability to zero.

To show that the last term in (A.2) goes to zero, consider the class of functions $\mathcal{F} \equiv \{f_d(x,z) \equiv (z - 1/4)1(x \leq d) | d \in [0,1]\}$ with the envelope $F(x,z) = 1$. The class $\mathcal{F}$ is formed by multiplying a fixed function $z \mapsto (z - 1/4)$ with a bounded Vapnik-Chervonenkis classes of functions $\{1(x \leq d) : 0 \leq d \leq 1\}$ and therefore satisfies the entropy condition in the third display on page 168 of van der Vaart and Wellner (1996). It follows that $\mathcal{F}$ satisfies the conditions of Theorem 2.8.1 of van der Vaart and Wellner (1996) and is therefore uniformly Glivenko–Cantelli for the class of probability measures $\{P_m\}$, i.e.,

$$\sup_{m \geq 1} P_m(\sup_{n \geq k} \|\mathbb{M}_{m,n}(\cdot, \tau_0) - M_m(\cdot)\|_\infty > \epsilon) \to 0$$

for every $\epsilon > 0$ as $k \to \infty$. Thus, we get $P(|\hat{d}_{m,n}(\hat{\tau}) - d_m| > \epsilon) \to 0$ as $m, n \to \infty$. This completes the proof of the theorem. $\square$

## A.2 Proof of Theorem 2.2

Recall that $T(x, \tau_0) = \sqrt{nh_n}(\hat{\mu}(x) - \tau_0)$. The following standard result from non-parametric regression theory is useful in proving Theorem 2.2. The proof follows, for example, from the results in Section 2.2 of Bierens (1987).

**Lemma A.3.** *Assume that $\mu(\cdot)$ and $\sigma^2(\cdot)$ is continuous on $[0,1]$. is continuous on [0,1]. We then have:*

*(i) For $0 < x, y < d_0$ and $x \neq y$,*

$$\begin{pmatrix} T(x,\tau_0) \\ T(y,\tau_0) \end{pmatrix} \to N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \bar{K}^2\sigma^2(x)/(mf(x)) & 0 \\ 0 & \bar{K}^2\sigma^2(y)/(mf(y)) \end{pmatrix} \right),$$

*in distribution.*

*(ii) For $d_0 < z < 1$, $T(z, \tau_0) \to \infty$ in probability.*

We now prove Theorem 2.2. Let $\nu(x)$ and $M(d)$ be as defined in proof of Theorem 2.1, e.g., $\nu(x) = (1/2)1(x \leq d_0)$. For notational convenience, let $Z_i(\tau_0) = \tilde{p}_n(X_i) = 1 - \Phi(T(X_i, \tau_0))$. We eventually show that $\|\mathbb{M}_n(\cdot, \hat{\tau}) - M(\cdot)\|_\infty$ converges to 0 in probability and then apply argmax continuous mapping theorem to prove consistency. By calculations similar to those in the proof of Theorem 2.1, $\|\mathbb{M}_n(\cdot, \hat{\tau}) - \mathbb{M}_n(\cdot, \tau_0)\| \leq \{2\Phi\left(\sqrt{nh_n}|\hat{\tau} - \tau_0|/2\right) - 1\}$, which converges to 0 in probability. So, it suffices to show that $\|\mathbb{M}_n(\cdot, \tau_0) - M(\cdot)\|_\infty$ converges to 0 in probability. We first establish marginal convergence. We have

$$
\begin{aligned}
& E\left[\Phi(T(X_1, \tau_0)) | X_1 = x\right] \qquad\qquad (A.3) \\
&= E\left[\Phi\left(\frac{(nh_n)^{-1/2}\left[(\mu(x) - \tau_0 + \epsilon_1)K(0) + \sum_{i=2}^n (Y_i - \tau_0)K\left(h_n^{-1}(x - X_i)\right)\right]}{(nh_n)^{-1}\left[K(0) + \sum_{i=2}^n K\left(h_n^{-1}(x - X_i)\right)\right]}\right)\right].
\end{aligned}
$$

The first term, both in the numerator and the denominator of the argument, is asymptotically negligible and thus, the expression in (A.3) equals $E[\Phi(T(x, \tau_0) + o_p(1))]$. Using Lemma A.3, this converges to $1 - \nu(x)$, by definition of weak convergence. As $Z_i(\tau_0) = 1 - \Phi(T(X_i, \tau_0))$, we get $E\left[\mathbb{M}_n(d, \tau_0)\right] = E[E\left\{Z_1(\tau_0) - 0.25\right\}1(X_1 \leq d)|X_1]$ which converges to $M(d)$. Further, $\mathrm{var}(\mathbb{M}_n(d, \tau_0)) = n^{-1}\mathrm{var}\left[\{Z_1(\tau_0) - 0.25\}1(X_1 \leq d)\right] + n^{-1}(n-1)\mathrm{cov}\left[\{Z_1(\tau_0) - 0.25\}1(X_1 \leq d), \{Z_2(\tau_0) - 0.25\}1(X_2 \leq d)\right]$. The first term in this expression goes to zero as $|Z_1(\tau_0)| \leq 1$. For $y \neq x$, by calculations similar to (A.3), $E\left[Z_1(\tau_0)Z_2(\tau_0) | X_1 = x, X_2 = y\right] = E\left[\Phi\left(T(x, \tau_0) + o_p(1)\right)\Phi\left(T(y, \tau_0) + o_p(1)\right)\right]$. Using Lemma A.3, $T(x, \tau_0)$ and $T(y, \tau_0)$ are asymptotically independent. Thus, by taking iterated expectations, it can be shown that

$$
\mathrm{cov}\left[\{Z_1(\tau_0) - 0.25\}1(X_1 \leq d), \{Z_2(\tau_0) - 0.25\}1(X_2 \leq d)\right]
$$

converges to 0. This justifies pointwise convergence, e.g., $\mathbb{M}_n(d, \hat{\tau}_0) - M(d) = o_p(1)$, for $d \in [0,1]$. Further, as $|Z_i(\hat{\tau}) - 1/4| \le 1$, for $d_1 < d < d_2$, we have

$$
\begin{aligned}
& E\left[|\{\mathbb{M}_n(d,\tau_0) - \mathbb{M}_n(d_1,\tau_0)\}\{\mathbb{M}_n(d_2,\tau_0) - \mathbb{M}_n(d,\tau_0)\}|\right] \\
& \qquad \le E\left[\left\{\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in (d_1, d])\right\}\left\{\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in (d, d_2])\right\}\right] \\
& \qquad = E\left[\frac{1}{n^2}\sum_{i\neq j} 1(X_i \in (d_1, d]) 1(X_j \in (d, d_2])\right] \\
& \qquad = \frac{1}{n^2}\sum_{i\neq j} F((d_1, d])F((d, d_2]) \\
& \qquad = \frac{n(n-1)}{n^2} F((d_1, d])F((d, d_2]).
\end{aligned}
$$

Note that the terms in the sum on the right side with $i = j$ are zero as $(d_1, d]$ and $(d, d_2]$ are disjoint. Further, the expression on the right side is bounded by $\|f\|_\infty^2 (d-d_1)(d_2-d) \le \|f\|_\infty^2 (d_2-d_1)^2$. As $f$ is continuous on $[0,1]$, $\|f\|_\infty < \infty$. Thus, the processes $\{\mathbb{M}_n(\cdot, \tau_0)\}_{n\ge 1}$ are tight in $D[0,1]$ using Theorem 15.6 in Billingsley (1968). So, $\mathbb{M}_n(\cdot, \tau_0)$ converges weakly to $M$ as processes in $D[0,1]$. As the limiting process is degenerate and the map $x(\cdot) \mapsto \sup_{d\in[0,1]} |x(d)|$ is continuous, by continuous mapping, we get $\|\mathbb{M}_n(\cdot, \tau_0) - M(\cdot)\|$ converges in probability to zero. As $d_0$ is the unique maximizer of the continuous function $M(\cdot)$ and $\hat{d}_n(\hat{\tau})$ is tight as $\hat{d}_n(\hat{\tau}) \in [0,1]$. Hence, by argmax continuous mapping theorem in van der Vaart and Wellner (1996), we get the result. □

## A.3 Proof of Theorem 2.5

We start with some notation. Let $W_m^{(i)} \equiv \sqrt{m}\bar{\epsilon}_{i,\cdot}/\sigma_0$ $(i = 1, \ldots, n)$, and consider our data as: $\{X_i, W_m^{(i)}\}(i = 1, \ldots, n)$. The variable $W_m^{(i)}$ has density $\phi_m(\cdot)$. Let $\mathbb{P}_{n,m}(\cdot)$ denote the empirical measure of these observables and $P_m$ the joint law of $\{X_1, W_m^{(1)}\}$. Let $\sigma_0$ denote the true variance of $\epsilon_{ij}$, and let $\sigma$ denote any such generic value. For a fixed $\sigma > 0$ and $h \in \mathbb{R}$ define, with a slight abuse of notation, $Z_{im}^{\sigma}(h) = 1 - \Phi\left((\sqrt{m}(\overline{Y}_{i\cdot} - \tau_0) - h)/\sigma\right) = 1 - \Phi\left([\sqrt{m}(\mu(X_i) - \tau_0) - h + \sqrt{m}\bar{\epsilon}_{i,\cdot}]/\sigma\right), \mathbb{M}_{n,m}^{\sigma}(h) = n^{-1}\sum_{i=1}^{n}\left\{Z_{im}^{\sigma}(h) - 1/2\right\}^2 = \mathbb{P}_{n,m}\left\{Z_{1m}^{\sigma}(h) - 1/2\right\}^2$, and note that

$$\hat{h}_{m,n}^{\hat{\sigma}} = \arg\min_h \mathbb{M}_{n,m}^{\hat{\sigma}}(h) \equiv \sqrt{m}(\hat{\tau}_{m,n}^{\hat{\sigma}} - \tau_0),$$

where $\hat{\sigma} = \hat{\sigma}_{m,n}$. Let $h_m^{\sigma} = \arg\min_h M_m^{\sigma}(h)$ where

$$\begin{aligned} M_m^{\sigma}(h) &= P_m\left[1/2 - \Phi\left[[\sqrt{m}(\mu(X_1) - \tau_0) - h + \sigma_0 W_m^{(1)}]/\sigma\right]\right]^2 \\ &= \int_0^1 \left[\int_{-\infty}^{\infty}\left\{1/2 - g_m^{\sigma,h}(x,y)\right\}^2 \phi_m(y)dy\right] p_X(x)dx, \end{aligned}$$

with $g_m^{\sigma,h}(x,y) = \Phi\left[[\sqrt{m}(\mu(x) - \tau_0) - h + \sigma_0 y]/\sigma\right]$.

Let $\epsilon, \xi > 0$ be given. We want to show that $P(|\hat{h}_{m,n}^{\hat{\sigma}} - 0| > \epsilon) \leq \xi$ for all large $m$ and $n$.We bound the quantity of interest as

$$P(|\hat{h}_{m,n}^{\hat{\sigma}} - 0| > \epsilon) \leq P(|\hat{h}_{m,n}^{\hat{\sigma}} - h_m^{\hat{\sigma}}| > \epsilon/2) + P(|h_m^{\hat{\sigma}} - 0| > \epsilon/2). \qquad \text{(A.4)}$$

We employ the following steps to complete the proof of the theorem:

*Step 1*: Establish that there exists $\delta_0 > 0$ and $M_0 > 0$ such that $|\sigma - \sigma_0| \leq \delta_0$ and $m \geq M_0$ implies $|h_m^{\sigma} - 0| < \epsilon/2$. Notice that as $\hat{\sigma}$ is a consistent estimator of $\sigma_0$, there exists $M_1$ such that for all $m, n \geq M_1 > 0$, $P(|\hat{\sigma}_{m,n} - \sigma_0| \leq \delta_0) \geq 1 - \xi/3$. Therefore, using *Step 1*, $P(|h_m^{\hat{\sigma}} - 0| > \epsilon/2) \leq \xi/3$ for $m \geq \max(M_0, M_1)$.

*Step 2*: The first term on the right side in (A.4) is bounded by $P(|\hat{h}_{m,n}^{\hat{\sigma}} - h_m^{\hat{\sigma}}| > \epsilon/2, |\hat{\sigma} - \sigma_0| \leq \delta_0) + P\left(|\hat{\sigma} - \sigma_0| > \delta_0\right) \leq P\left(\sup_{|\sigma - \sigma_0| \leq \delta_0}|\hat{h}_{m,n}^{\sigma} - h_m^{\sigma}| > \epsilon/2\right) + \xi/3$ for all $n, m \geq \max(M_0, M_1)$. Therefore, it is enough to show that for some $M$, possibly depending on $\epsilon$,

$$\sup_{m\geq M} P\left(\sup_{|\sigma-\sigma_0|\leq\delta_0} |\hat{h}^{\sigma}_{m,n} - h^{\sigma}_{m}| > \epsilon/2\right) \to 0, \text{ as } n\to\infty. \tag{A.5}$$

*Proof of Step 1*: We study the behavior of $M^{\sigma}_m(h)$ as $m\to\infty$. Note that $g^{\sigma,h}_m(x,y)\to \Phi\left((-h+\sigma_0 y)/\sigma\right)$, if $x\leq d^0$, and 1 if $x>d^0$, as $m\to\infty$. Therefore, $M^{\sigma}_m(h)$ converges point–wise, by the dominated convergence theorem along with Scheffe's theorem, to $M^{\sigma}(h)$, where

$$M^{\sigma}(h) = c^{\sigma}_1(h)\int_0^{d^0} p_X(x)dx + \frac{1}{4}\int_{d^0}^{1} p_X(x)dx < 1/4, \tag{A.6}$$

with $c^{\sigma}_1(h) = \int_{-\infty}^{\infty}\left[1/2-\Phi\left((-h+\sigma_0 y)/\sigma\right)\right]^2\phi(y)dy$. To see this, observe that $\int_{-\infty}^{\infty}\left\{1/2-g^{\sigma,h}_m(x,y)\right\}^2\phi_m(y)dy$, which is uniformly bounded by a positive constant for all $m$ and $x$, can be decomposed as

$$\int_{-\infty}^{\infty}\left\{1/2-g^{\sigma,h}_m(x,y)\right\}^2\phi(y)dy + \int_{-\infty}^{\infty}\left\{1/2-g^{\sigma,h}_m(x,y)\right\}^2\{\phi_m(y)-\phi(y)\}dy,$$

where the first term converges to $c_1(h)$ for $x\leq d^0$ and to $1/4$ for $x>d^0$. The second term converges to 0 by Scheffe's theorem for all $x\in[0,1]$. The convergence of $M^{\sigma}_m(h)$ now directly follows from the dominated convergence theorem. Let $h^{\sigma} = \arg\min M^{\sigma}(h)$ for $h\in\mathbb{R}$.

We claim that there exists $\delta'>0$, such that $\sup_{|\sigma-\sigma_0|\leq\delta'}\sup_{h\in\mathbb{R}}|M^{\sigma}_m(h)-M^{\sigma}(h)|\to 0$ as $m\to\infty$. In course of justifying this claim, we will write $\Phi(x)\left(1-\Phi(x)\right)$ as $\Phi\left(1-\Phi\right)(x)$ for notational convenience. Choose $\delta'$ such that $0<\delta'<\sigma_0$. Let $\eta>0$ be given. Note that $M^{\sigma}_m(h)-M^{\sigma}(h) = A^{\sigma,h}_m\int_0^{d^0} p_X(x)dx + \int_{d^0}^{1} B^{\sigma,h}_m(x)p_X(x)dx$, where

$$A^{\sigma,h}_m = \int_{-\infty}^{\infty}\left[1/2-\Phi\left((-h+\sigma_0 y)/\sigma\right)\right]^2(\phi_m-\phi)(y)dy \text{ and}$$

$$B_m^{\sigma,h}(x) = \int_{-\infty}^{\infty} \left\{1/2 - g_m^{\sigma,h}(x,y)\right\}^2 \phi_m(y)dy - 1/4.$$

To simplify notation, denote the set $\{(\sigma,h) : |\sigma - \sigma_0| \leq \delta', h \in \mathbb{R}\}$ by $\mathcal{C}$. Then,

$$\sup_{\mathcal{C}} |M_m^\sigma(h) - M^\sigma(h)| \quad \leq \quad F_X(d^0) \sup_{\mathcal{C}} |A_m^{\sigma,h}| + \sup_{\mathcal{C}} \int_{d^0}^{1} |B_m^{\sigma,h}(x)|\, p_X(x)\, dx\,.$$

Now, $\sup_{\mathcal{C}} |A_m^{\sigma,h}| \leq \int_{-\infty}^{\infty} |\phi_m - \phi|(y)dy \to 0$ by Scheffe's theorem, and

$$\begin{aligned} |B_m^{\sigma,h}(x)| \quad &\leq \quad \int_{-\infty}^{\infty} |\phi_m - \phi|(y)dy + \sup_{\mathcal{C}} \left| \int_{-\infty}^{\infty} \left\{1/2 - g_m^{\sigma,h}(x,y)\right\}^2 \phi(y)dy - 1/4 \right| \\ &= \quad o(1) + \sup_{\mathcal{C}} \int_{-\infty}^{\infty} \Phi\,(1-\Phi)\left((\sqrt{m}(\mu(x) - \tau_0) - h + \sigma_0 y)/\sigma\right) \phi(y)dy\,. \end{aligned}$$

Also,

$$\begin{aligned} &\sup_{\mathcal{C}} \int_{d^0}^{1} |B_m^{\sigma,h}(x)|\, p_X(x)\, dx \\ &\quad = \quad \left\{ \sup_{\mathcal{C}_{\leq 0}} \int_{d^0}^{1} |B_m^{\sigma,h}(x)|\, p_X(x)\, dx \right\} \vee \left\{ \sup_{\mathcal{C}_{>0}} \int_{d^0}^{1} |B_m^{\sigma,h}(x)|\, p_X(x)\, dx \right\}, \end{aligned}$$

where $\mathcal{C}_{\leq 0}$ and $\mathcal{C}_{>0}$ are defined analogously to $\mathcal{C}$, but with $h$ varying over $(-\infty, 0]$ and $(0,\infty)$, respectively. For each $x > d^0$,

$$\sup_{\mathcal{C}_{\leq 0}} \int_{-\infty}^{\infty} \Phi\,(1-\Phi)\left((\sqrt{m}(\mu(x) - \tau_0) - h + \sigma_0 y)/\sigma\right) \phi(y)dy$$

is easily seen to be dominated by

$$\sup_{|\sigma - \sigma_0| \leq \delta'} \int_{-\infty}^{\infty} (1-\Phi)\left[[\sqrt{m}(\mu(x) - \tau_0) + \sigma_0 y]/\sigma\right] \phi(y)dy,$$

which goes to 0 as $m \to \infty$. It follows readily that the first term on the right side of the last display is $o(1)$. It remains to deal with the second. To this end, for $\lambda, h > 0$, define $D_m^{\lambda,h} = \{d^0 < x \le 1 : |\mu(x) - (\tau_0 + h/\sqrt{m})| \le \lambda\}$. Given $\eta > 0$, there exists $\lambda \equiv \lambda(\eta) > 0$, not depending on $h > 0$, such that $\int_{D_m^{\lambda,h}} p_X(x)\, dx < \eta$ by Assumption (A) of Theorem 3 in the paper. Then,

$$\begin{aligned}
&\sup_{\mathcal{C}_{>0}} \left| \int_{d^0}^{1} B_m^{\sigma,h}(x) p_X(x) dx \right| \\
&\le \sup_{\mathcal{C}} \left| \int_{D_m^{\lambda,h}} B_m^{\sigma,h}(x) p_X(x) dx \right| + \sup_{\mathcal{C}_{>0}} \left| \int_{[d^0,1]-D_m^{\lambda,h}} B_m^{\sigma,h}(x) p_X(x) dx \right| \\
&\le \eta + o(1) \\
&\quad + \sup_{\mathcal{C}_{>0}} \int_{[d^0,1]-D_m^{\lambda,h}} \int_{-\infty}^{\infty} \Phi(1-\Phi) \left[ \frac{\sqrt{m}(\mu(x) - \tau_0 - h m^{-1/2}) + \sigma_0 y}{\sigma} \right] \phi(y) dy\, p_X(x) dx\,.
\end{aligned}$$

The last term in the above display is readily seen to be bounded by

$$\int_{d^0}^{1} \int_{-\infty}^{\infty} \sup_{|\sigma-\sigma_0| \le \delta'} \max\left\{ \Phi\left( \frac{-\sqrt{m}\lambda + \sigma_0 y}{\sigma} \right), (1-\Phi)\left( \frac{\sqrt{m}\lambda + \sigma_0 y}{\sigma} \right) \right\} \phi(y) dy\, p_X(x) dx$$

which can be made less than $\eta$ for large $m$. It follows that $\sup_{\mathcal{C}_{>0}} \int_{d^0}^{1} |B_m^{\sigma,h}(x)|\, p_X(x)\, dx < 3\,\eta$ for all sufficiently large $m$ and the claim follows.

Next, we claim that there exists there exists $\delta_0 > 0$ and $M_0 > 0$ such that for all $\sigma$ with $|\sigma - \sigma_0| \le \delta_0$ and $m \ge M_0$, $|h_m^{\sigma} - 0| < \epsilon/2$. This is proved by a direct application of Theorem A.In that theorem, take $n$ to be $m$, $\mathcal{T}$ to be the set $|\sigma - \sigma_0| \le \delta'$ and $\mathcal{H}$ to be $\mathbb{R}$. Also, $\mathbb{M}_n^{\tau}$ is now $-M_m^{\sigma}$ and $M^{\tau}$ is now $-M^{\sigma}$. We will show that $-M^{\sigma}$ is uniquely maximized at a point, say $h^{\sigma}$, and also that $\inf_{|\sigma-\sigma_0|\le\delta'} \inf_{|h-h^{\sigma}|>\eta} \{M^{\sigma}(h) - M^{\sigma}(h^{\sigma})\} > 0$ for every $\eta > 0$, whence, by by the previous claim, it will follow that $\sup_{|\sigma-\sigma_0|\le\delta'} |h_m^{\sigma} - h^{\sigma}|$ converges to 0 with increasing $m$. But, as will also be seen, $h^{\sigma}$ equals 0 for all $\sigma$ and hence the claim follows with

$\delta_0$ taken to be $\delta'$.

From the form of $M^\sigma(h)$ (see (A.6)) it suffices to show that

$$\inf_{|\sigma-\sigma_0|\le\delta'}\ \inf_{|h-h^\sigma|>\eta}\{c_1^\sigma(h)-c_1^\sigma(h^\sigma)\}>0,$$

where $h^\sigma$ is the unique point at which $c_1^\sigma$ is minimized. We now make some change of variables to facilitate the ensuing argument. Define $\lambda=\sigma/\sigma_0$ and $s=h/\sigma_0$. Then $|\sigma-\sigma_0|\le\delta'\Leftrightarrow|\lambda-1|\le\delta''$ (for some $\delta''<1$) and $\Phi((-h+\sigma_0\,y)/\sigma)=\Phi(\lambda^{-1}(y-s))$. Defining $\tilde{c}_1^\lambda(s)=\int_{-\infty}^{\infty}\left[1/2-\phi(\lambda^{-1}(y-s))\right]^2\phi(y)\,dy$, it suffices to show that $\inf_{|\lambda-1|\le\delta''}\inf_{|s-s^\lambda|\ge\eta/\sigma_0}\{\tilde{c}_1^\lambda(s)-\tilde{c}_1^\lambda(s_\lambda)\}>0$ where $s^\lambda$ is the unique minimizer of $\tilde{c}_1^\lambda$. It is easy to see that $\tilde{c}_1^\lambda(s)=E\left[1/2-\Phi(\lambda^{-1}(Z-s))\right]^2$ where $Z$ is a standard normal random variable. By the symmetry of $Z$ about 0, it follows easily that $\tilde{c}_1^\lambda(s)=\tilde{c}_1^\lambda(-s)$. Furthermore $\tilde{c}_1^\lambda(s)$ is strictly increasing for $s>0$, and is therefore strictly decreasing for $s\le0$, showing that 0 is the unique minimizer of $\tilde{c}_1^\lambda$. Hence $s_\lambda=0$ for all $\lambda$, showing that $h^\sigma=0$ for all $\sigma$. Thus,

$$\inf_{|\lambda-1|\le\delta''}\ \inf_{|s-s^\lambda|\ge\eta/\sigma_0}\{\tilde{c}_1^\lambda(s)-\tilde{c}_1^\lambda(s_\lambda)\}=\inf_{|\lambda-1|\le\delta''}\{\tilde{c}_1^\lambda(\eta/\sigma_0)-\tilde{c}_1^\lambda(0)\}\,.$$

Since $\tilde{c}_1^\lambda(\eta/\sigma_0)-\tilde{c}_1^\lambda(0)$ is continuous and positive for each $\lambda$, its infimum on the set $|\lambda-1|\le\delta''$, which must be achieved, is positive.

*Proof of Step 2*: Consider the class of functions $\mathcal{F}_\infty\equiv\cup_m\mathcal{F}_m$ where $\mathcal{F}_m\equiv\{f_{h,\sigma}(x,w)\equiv[1/2-\Phi(\sqrt{m}(\mu(x)-\tau_0)/\sigma+h/\sigma+w\sigma_0/\sigma)]^2|\,\tau\in\mathbb{R},\sigma\in[\sigma_0-\delta_0,\sigma_0+\delta_0]\}$. This is a subclass of the large class of functions $\mathcal{G}=\{g_{\alpha,\beta,\gamma}(x,w)\equiv[1/2-\Phi(\alpha\mu(x)+\beta w+\gamma)]^2|\,(\alpha,\beta,\gamma)\in\mathbb{R}^3\}$. The class $\{\alpha\mu(x)+\beta w+\gamma\}$ as $(\alpha,\beta,\gamma)$ varies in $\mathbb{R}^3$ forms a finite dimensional vector space of measurable functions and is therefore a Vapnik–Chervonenkis class. Hence, $1/2-\Phi(\alpha\mu(x)+\beta w+\gamma)$, being their bounded monotone transformation, is a bounded Vapnik–Chervonenkis class and consequently, so is $\mathcal{F}_\infty$. Thus, $\mathcal{F}_\infty$ satisfies the entropy condition in the third display on Page 168 of van der Vaart and Wellner (1996) and therefore the conditions of Theorem 2·8·1 of van der Vaart and Wellner (1996) and is uniformly Glivenko–Cantelli for the class of probability measures $\{P_m\}$, i.e., for any given $\zeta > 0$, $\sup_{m\geq 1} P_m(\sup_{k\geq n} \|\mathbb{M}^{\sigma}_{m,k} - M^{\sigma}_{m}\|_{\mathcal{F}_\infty > \zeta}) \to 0$ as $n \to \infty$ and therefore,

$$\sup_{m\geq 1} P_m\left(\sup_{k\geq n} \|\mathbb{M}^{\sigma}_{m,k} - M^{\sigma}_{m}\|_{\mathcal{F}_m} > \zeta\right) \to 0 \ \text{ as } n \to \infty. \tag{A.7}$$

Next, using *generic steps*, we can show that $\sup_{|\sigma-\sigma_0|\leq\delta_0} |\hat{h}^{\sigma}_{m,n} - h^{\sigma}_{m}| > \epsilon/2 \Rightarrow \|\mathbb{M}^{\sigma}_{m,n} - M^{\sigma}_{m}\|_{\mathcal{F}_m} \geq \eta_m(\epsilon/2)$ where $\eta_m(\epsilon) = \inf_{|\sigma-\sigma_0|\leq\delta_0} \inf_{|h-h^{\sigma}_{m}|>\epsilon/2}\{M^{\sigma}_{m}(h) - M^{\sigma}_{m}(h^{\sigma}_{m})\}$.

Next, we claim that there exists $\eta > 0$ and an integer $\tilde{M}$ such that $\eta_m(\epsilon) \geq \eta > 0$ for all $m \geq \tilde{M}$. To see this, note that by the previous claim, for all sufficiently large $m$, uniformly for $\sigma \in [\sigma_0 - \delta_0, \sigma_0 + \delta_0]$, we have $[h^{\sigma}_{m} - \epsilon/2, h^{\sigma}_{m} + \epsilon/2]^c \subset [-\epsilon/4, \epsilon/4]^c$. We conclude, that for all sufficiently large $m$,

$$\eta_m(\epsilon) \geq \tilde{\eta}_m(\epsilon) \equiv \inf_{|\sigma-\sigma_0|\leq\delta_0} \inf_{|h-0|>\epsilon/4}\{M^{\sigma}_{m}(h) - M^{\sigma}_{m}(h^{\sigma}_{m})\}.$$

For $h$ and $\sigma$ such that $|h-0| > \epsilon/4$ and $|\sigma-\sigma_0| \leq \delta_0$, we can bound $M^{\sigma}_{m}(h) - M^{\sigma}_{m}(h^{\sigma}_{m}) = (M^{\sigma}_{m} - M^{\sigma})(h) - M^{\sigma}_{m}(h^{\sigma}_{m}) + M^{\sigma}(h) \geq -\sup_{|h-0|>\epsilon/4} |(M^{\sigma}_{m} - M^{\sigma})(h)| - M^{\sigma}_{m}(0) + M^{\sigma}(h)$ further from below by

$$\begin{aligned} -\sup_{|\sigma-\sigma_0|\leq\delta_0} \sup_{|h|>\epsilon/4} |(M^{\sigma}_{m} - M^{\sigma})(h)| - \sup_{|\sigma-\sigma_0|\leq\delta_0} |(M^{\sigma}_{m} - M^{\sigma})(0)| \\ + \inf_{|\sigma-\sigma_0|\leq\delta_0} \inf_{|h|>\epsilon/4}\{M^{\sigma}(h) - M^{\sigma}(0)\}. \end{aligned}$$

As $\sup_{|\sigma-\sigma_0|\leq\delta_0} \sup_{|h-0|>\epsilon/4} |(M^{\sigma}_{m} - M^{\sigma})(h)| \to 0$ and $\sup_{|\sigma-\sigma_0|\leq\delta_0} |(M^{\sigma}_{m} - M^{\sigma})(0)| \to 0$ as $m \to \infty$, and $\eta = \inf_{|\sigma-\sigma_0|\leq\delta_0} \inf_{|h-0|>\epsilon/4}\{M^{\sigma}(h) - M^{\sigma}(0)\}/2 > 0$, it follows that for all large $m$, $\tilde{\eta}_m(\epsilon) \geq \eta > 0$; therefore, for all sufficiently large $m$, say $m \geq \tilde{M}$, $\eta_m(\epsilon) \geq \eta > 0$. This completes the proof of the claim.

Hence, for all $m \geq \tilde{M}$,

$$
\begin{aligned}
& \sup_{m \geq \tilde{M}} P_m \left( \sup_{|\sigma - \sigma_0| \leq \delta_0} |\hat{h}^{\sigma}_{m,n} - h^{\sigma}_{m}| > \epsilon/2 \right) \\
& \leq \sup_{m \geq \tilde{M}} P_m \left\{ \sup_{k \geq n} \|\mathbb{M}^{\sigma}_{m,n} - M^{\sigma}_{m}\|_{\mathcal{F}_m} > \eta_m(\epsilon)/2 \right\} \\
& \leq \sup_{m \geq \tilde{M}} P_m \left( \sup_{k \geq n} \|\mathbb{M}^{\sigma}_{m,k} - M^{\sigma}_{m}\|_{\mathcal{F}_m} > \eta/2 \right)
\end{aligned}
$$

and (A.5) follows from (A.7). $\square$

# APPENDIX B

# Proofs for Chapter 3

We first state a result which is useful in deriving the rate of convergence of our estimators. As earlier, we readily use the notations '$\lesssim$' and '$\gtrsim$' to imply that the corresponding inequalities ($<$ and $>$) hold up to some positive constant multiple. We use $E^*$ to denote the outer expectation with respect to the concerned probability measure.

**Theorem B.1.** *Let $\{\mathbb{M}_n(d,\sigma),\ n \geq 1\}$ be stochastic processes and $\{M_n(d,\sigma),\ n \geq 1\}$ be deterministic functions, indexed by $d \in \Theta$ and $\sigma \in \Sigma$. Let $d_n \in \Theta$, $\sigma_0 \in \Sigma$ and $\kappa > 0$ be arbitrary, and $d \mapsto \rho_n(d, d_n)$ be an arbitrary map from $\Theta$ to $[0,\infty)$. Let $\hat{d}_n$ be a point of minimum of $\mathbb{M}_n(d,\hat{\sigma}_n)$, where $\hat{\sigma}_n$ is random. For each $\epsilon > 0$, suppose that the following hold:*

*(a) There exists a sequence of sets $U_{n,\epsilon}$ in $\Sigma$ which contain $\sigma_0$ and $P[\hat{\sigma}_n \notin U_{n,\epsilon}] < \epsilon$.*

*(b) For all sufficiently large $n$, $0 < \delta < \kappa$, and $d$ such that $\rho_n(d, d_n) < \kappa$,*

$$M_n(d,\sigma_0) - M_n(d_n,\sigma_0) \gtrsim \rho_n^2(d,d_n),$$

$$E^* \sup_{\substack{\rho_n(d,d_n)<\delta \\ \sigma \in U_{n,\epsilon}}} |(\mathbb{M}_n(d,\sigma) - M_n(d,\sigma_0)) - (\mathbb{M}_n(d_n,\sigma) - M_n(d_n,\sigma_0))| \quad \leq \quad C_\epsilon \frac{\phi_n(\delta)}{\sqrt{n}},$$

*for a constant $C_\epsilon > 0$ and functions $\phi_n$ (not depending on $\epsilon$) such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$.*

*Suppose that $r_n$ satisfies*

$$r_n^2 \ \phi_n\left(\frac{1}{r_n}\right) \lesssim \sqrt{n},$$

*and $\rho_n(\hat{d}_n, d_n)$ converges to zero in probability; then $r_n \ \rho_n(\hat{d}_n, d_n) = O_P(1)$.*

This theorem puts together the results in Theorem 3.2.5 in van der Vaart and Wellner (1996) and Theorem 5.2 in Banerjee and McKeague (2007).

## B.1 Proof of Theorem 3.2

The following lemma gives the explicit distance function $\rho_n$ that is used in proving Theorem 3.2.

**Lemma B.2.** *Fix $\eta > 0$. Let the map $d \mapsto \rho_n^2(d, d_{m,n})$ from $(0,1)$ to $[0,\infty)$ be*

$$K_1\left[|d-d_0|\,1\,(d<d_0) + \left|d - d_{m,n} - \frac{\eta}{m^{1/(2k)}}\right| 1\left(d > d_{m,n} + \frac{\eta}{m^{1/(2k)}}\right)\right], \quad \text{(B.1)}$$

*for some $K_1 > 0$. Then $K_1$ and $\kappa > 0$ can be chosen such that for sufficiently large $n$ and $\rho_n(d, d_{m,n}) < \kappa$, we have*

$$M_{m,n}(d) - M_{m,n}(d_{m,n}) \geq \rho_n^2(d, d_{m,n}).$$

Using this lemma, we first give a proof of Theorem 3.2. Note that $\sqrt{mn}(\hat{\sigma} - \sigma_0) = O_P(1)$. So, given $\epsilon > 0$, there exists $L_\epsilon > 0$ such that $P[\sqrt{mn}|\hat{\sigma} - \sigma_0| \leq L_\epsilon] > 1 - \epsilon$. Let $U_{n,\epsilon} = [\sigma_1, \sigma_2] = [\sigma_0 - L_\epsilon/\sqrt{mn}, \sigma_0 + L_\epsilon/\sqrt{mn}]$ and let $\mathbb{G}_n$ denote the empirical process, i.e., $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_n)$. For $\kappa$ as in Lemma B.2, $0 \leq \delta < \kappa$, and $\rho_n$ as defined in (B.1), consider the expression

$$
\begin{aligned}
&E^* \sup_{\substack{\rho_n(d,d_{m,n})<\delta \\ \sigma\in U_{n,\epsilon}}} \sqrt{n}|(\mathbb{M}_{m,n}(d,\sigma) - M_{m,n}(d,\sigma_0)) - (\mathbb{M}_{m,n}(d_{m,n},\sigma) - M_{m,n}(d_{m,n},\sigma_0))| \\
&\leq E^* \sup_{\substack{\rho_n(d,d_{m,n})<\delta \\ \sigma\in U_{n,\epsilon}}} \sqrt{n}|(\mathbb{M}_{m,n}(d,\sigma) - \mathbb{M}_{m,n}(d_{m,n},\sigma)) - (M_{m,n}(d,\sigma) - M_{m,n}(d_{m,n},\sigma))| \\
&+ \sup_{\substack{\rho_n(d,d_{m,n})<\delta \\ \sigma\in U_{n,\epsilon}}} \sqrt{n}|(M_{m,n}(d,\sigma) - M_{m,n}(d_{m,n},\sigma)) - (M_{m,n}(d,\sigma_0) - M_{m,n}(d_{m,n},\sigma_0))| \\
&\leq E^* \sup_{\substack{|d-d_{m,n}|<\delta^2/K_1+Am^{-1/2k} \\ \sigma\in U_{n,\epsilon}}} \left|\mathbb{G}_n\left[\left(\Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma}\right) - \gamma\right)(1(X\leq d) - 1(X\leq d_{m,n}))\right]\right| \\
&+ \sqrt{n} \sup_{\substack{|d-d_{m,n}|<\delta^2/K_1+Am^{-1/2k} \\ \sigma\in U_{n,\epsilon}}} \left|P_n\left[\left\{\Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma}\right) - \Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma_0}\right)\right\}(1(X\leq d) - 1(X\leq d_{m,n}))\right]\right|.
\end{aligned}
$$

The first term in the above display involves an empirical process acting on a class of functions, say $\mathcal{F}$. This class $\mathcal{F}$ is a product of two VC classes, $\{(\Phi(\sqrt{m}\ \cdot/\sigma) - \gamma) : \sigma \in U_{n,\epsilon}\}$ and $\{1(\cdot \leq d) - 1(\cdot \leq d_{m,n}) : |d - d_{m,n}| < \delta^2/K_1 + Am^{-1/2k}\}$, each with VC-index at most 3. Also, an envelope for this class is given by

$$
G(x) = 1\left[x \in (d_{m,n} - \delta^2/K_1 - Am^{-1/(2k)}, d_{m,n} + \delta^2/K_1 + Am^{-1/(2k)})\right]
$$

with $(P_nG^2)^{1/2} \lesssim \sqrt{2(\delta^2/K_1 + Am^{-1/(2k)})}$. Hence, the uniform entropy integral for $\mathcal{F}$ is bounded by a constant which only depends upon the VC-indices, i.e., the quantity

$$
J(1,\mathcal{F}) = \sup_Q \int_0^1 \sqrt{1 + \log N_C(\epsilon\|G\|_{Q,2}, \mathcal{F}, L_2(Q))}d\epsilon
$$

is bounded, where $N_C(\cdot)$ denotes the covering number; see Theorems 9.3 and 9.15 of Kosorok (2008) for more details. Using Theorem 2.14.1 of van der Vaart and Wellner

(1996), we have

$$E^* \sup_{\substack{|d-d_{m,n}|<\delta^2/K_1+Am^{-1/2k} \\ \sigma\in U_{n,\epsilon}}} \left| \mathbb{G}_n \left[ \left( \Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma}\right) - \gamma \right) (1(X \le d) - 1(X \le d_{m,n})) \right] \right|$$

$$\le \; J(1,\mathcal{F})(P_n G^2)^{1/2} \;\lesssim\; \sqrt{2(\delta^2/K_1 + Am^{-1/(2k)})}.$$

Note that for $\sigma \in U_{n,\epsilon} = [\sigma_1, \sigma_2]$, we have

$$\left| \Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma}\right) - \Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma_0}\right) \right| \le \left| \Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma_1}\right) - \Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma_2}\right) \right|.$$

Hence, by using the fact that $\Phi$ is Lipschitz of order 1, for sufficiently large $n$, we get

$$\begin{aligned}
&\sqrt{n} \sup_{\substack{|d-d_{m,n}|<\delta^2/K_1+Am^{-1/2k} \\ \sigma\in U_{n,\epsilon}}} \left| P_n \left[ \left\{ \Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma}\right) - \Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma_0}\right) \right\} (1(X \le d) - 1(X \le d_{m,n})) \right] \right| \\
&\qquad \le \; \sqrt{n} P_n \left[ \left| \Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma_1}\right) - \Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma_2}\right) \right| |G(X)| \right] \\
&\qquad \le \; \sqrt{n} \left[ P_n \left| \Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma_1}\right) - \Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma_2}\right) \right|^2 \right]^{\frac{1}{2}} (P_n G^2)^{\frac{1}{2}} \\
&\qquad \lesssim \; \sqrt{nm} \frac{\sigma_2 - \sigma_1}{\sigma_2 \sigma_1} (E\bar{Y}^2)^{1/2} \sqrt{2(\delta^2/K_1 + Am^{-1/(2k)})} \\
&\qquad \lesssim \; \frac{4L_\epsilon}{\sigma_0^2} (E\bar{Y}^2)^{1/2} \sqrt{2(\delta^2/K_1 + Am^{-1/(2k)})}.
\end{aligned}$$

As $E(\bar{Y}^2) = (1/m)E\{\mu(X)\}^2 + \sigma_0^2$ is bounded, we have

$$E^* \sup_{\substack{\rho_n(d,d_{m,n})<\delta \\ \sigma\in U_{n,\epsilon}}} \sqrt{n} |(\mathbb{M}_{m,n}(d,\sigma) - M_{m,n}(d,\sigma_0)) - (\mathbb{M}_{m,n}(d_{m,n},\sigma) - M_{m,n}(d_{m,n},\sigma_0))| \le C_\epsilon \phi_n(\delta), \tag{B.2}$$

for some $C_\epsilon > 0$ and $\phi_n(\delta) = \sqrt{\delta^2 + m^{-1/(2k)}}$. Also, $\rho_n^2(d, d_{m,n}) \le K_1(|d - d_0| + |d_0 - d_{m,n} - \eta m^{-1/(2k)}|) \to 0$, if $|d - d_0| \to 0$. So, $\rho_n(\hat{d}_{m,n}, d_{m,n})$ converges in probability to zero by consistency of $\hat{d}_{m,n}$. Then by Theorem B.1, the rate of convergence, say $r_n$, satisfies

$$
\begin{aligned}
r_n^2 \phi\left(\frac{1}{r_n}\right) \lesssim \sqrt{n} \ &\Rightarrow r_n^2 + r_n^4 m^{-1/(2k)} \leq n \\
&\Rightarrow r_n^2 \lesssim n \wedge \sqrt{n^{1+\beta/(2k)}}.
\end{aligned}
\tag{B.3}
$$

With $\alpha = \min(1, \beta/(2k)) = \min(1, 1/2 + \beta/(4k), \beta/(2k))$, $r_n^2 = n^\alpha$ satisfies (B.3). As $m^{-1/(2k)} \lesssim n^{-\alpha}$, we also have $n^\alpha(d_{m,n} + \eta m^{-1/(2k)} - d_0) = O(1)$. So, $n^\alpha \rho_n^2(\hat{d}_{m,n}, d_{m,n}) = O_P(1) \Rightarrow n^\alpha(\hat{d}_{m,n} - d_0) = O_P(1)$. As $m_0 n^{1+\beta} = N$, we get the result. □

**Proof of Lemma B.2.** Let $\epsilon > 0$ be chosen such that $\mu$ is increasing on $(d_0, d_0 + \epsilon)$. Let $f_0 = \inf_{d:|d-d_0|<\epsilon} f(d) > 0$. For $d \in (d_0 - \epsilon, d_0 + \epsilon)$,

$$
\begin{aligned}
M_{m,n}(d) - M_{m,n}(d_{m,n}) \geq\ & 1(d < d_0) f_0 \left[ (\Phi_n(0) - \gamma)(d - d_0) + M_{m,n}(d_0) - M_{m,n}(d_{m,n}) \right] \\
& + 1(d \geq d_0) f_0 \int_{d_{m,n}}^{d} \left[ \Phi_n\left( \frac{\sqrt{m}\mu(x)}{\sqrt{2}\sigma_0} \right) - \gamma \right] dx \\
\geq\ & 1(d < d_0) f_0 \left[ |\Phi_n(0) - \gamma| \, |d - d_0| \right] \\
& + 1(d \geq d_0) f_0 \int_{d_{m,n}}^{d} \left[ \Phi_n\left( \frac{\sqrt{m}\mu(x)}{\sqrt{2}\sigma_0} \right) - \gamma \right] dx.
\end{aligned}
$$

Recall that from (3.7),

$$
d_{m,n} = d_0 + \left[ \frac{k!\sqrt{2}\sigma_0 \Phi^{-1}(\gamma)}{\mu^{(k)}(d_0+)} \right]^{1/k} m^{-1/(2k)} + o(m^{-1/(2k)}). \tag{B.4}
$$

Hence, for sufficiently large $n$, $d_{m,n} + \eta m^{-1/(2k)} < d_0 + \epsilon$. For such large $n$'s and $d \in (d_{m,n} + \eta m^{-1/(2k)}, d_0 + \epsilon)$, we have:

$$
M_{m,n}(d) - M_{m,n}(d_{m,n}) \geq f_0 \int_{d_{m,n}}^{d} \left[ \Phi_n\left( \frac{\sqrt{m}\mu(x)}{\sqrt{2}\sigma_0} \right) - \gamma \right] dx
$$

$$\geq f_0 \int\limits_{d_{m,n}+\eta m^{-1/(2k)}}^{d} \left[\Phi_n\left(\frac{\sqrt{m}\mu(x)}{\sqrt{2}\sigma_0}\right) - \gamma\right] dx$$

$$\geq f_0 \ (d - (d_{m,n} + \eta m^{-1/(2k)})) \left[\Phi_n\left(\frac{\sqrt{m}\mu(d_{m,n} + \eta m^{-1/(2k)})}{\sqrt{2}\sigma_0}\right) - \gamma\right]. \quad \text{(B.5)}$$

Next, we show that $[\Phi_n\left(\sqrt{m}\mu(d_{m,n} + \eta m^{-1/(2k)})/(\sqrt{2}\sigma_0)\right) - \gamma]$ is bounded away from zero. By Pólya's theorem, $\Phi_n$ converge uniformly to $\Phi$. So, for sufficiently large $n$,

$$\begin{aligned}
&\Phi_n\left(\frac{\sqrt{m}\mu(d_{m,n} + \eta m^{-1/(2k)})}{\sqrt{2}\sigma_0}\right) - \gamma \\
&= \Phi_n\left(\frac{\sqrt{m}\mu(d_{m,n} + \eta m^{-1/(2k)})}{\sqrt{2}\sigma_0}\right) - \Phi_n\left(\frac{\sqrt{m}\mu(d_{m,n})}{\sqrt{2}\sigma_0}\right) \\
&> \frac{1}{2}\left[\Phi\left(\frac{\sqrt{m}\mu(d_{m,n} + \eta m^{-1/(2k)})}{\sqrt{2}\sigma_0}\right) - \Phi\left(\frac{\sqrt{m}\mu(d_{m,n})}{\sqrt{2}\sigma_0}\right)\right].
\end{aligned}$$

As $\Phi\left(\sqrt{m}\mu(d_{m,n})/(\sqrt{2}\sigma_0)\right)$ converges to $\gamma \in (0,1)$, $\sqrt{m}\mu(d_{m,n})$ is $O(1)$. Hence, it suffices to show that the difference $\sqrt{m}\{\mu(d_{m,n} + \eta m^{-1/(2k)}) - \mu(d_{m,n})\}$ is bounded away from zero. With $\tilde{\zeta}_n$ being some point between $d_0$ and $d_{m,n} + \eta m^{-1/(2k)}$ and $\zeta_n$ as defined in (3.6), we have

$$\begin{aligned}
&\sqrt{m}\{\mu(d_{m,n} + \eta m^{-1/(2k)}) - \mu(d_{m,n})\} \\
&= \frac{\sqrt{m}}{k!}\{\mu^{(k)}(\tilde{\zeta}_n)(d_{m,n} + \eta m^{-1/(2k)} - d_0)^k - \mu^{(k)}(\zeta_n)(d_{m,n} - d_0)^k\} \\
&> \frac{\sqrt{m}\mu^{(k)}(\tilde{\zeta}_n)}{k!}[(d_{m,n} + \eta m^{-1/(2k)} - d_0)^k - (d_{m,n} - d_0)^k] \\
&\quad + \frac{\sqrt{m}}{k!}[\mu^{(k)}(\tilde{\zeta}_n) - \mu^{(k)}(\zeta_n)](d_{m,n} - d_0)^k \\
&> \frac{\mu^{(k)}(d_0+)\eta^k}{k!} + o(1).
\end{aligned}$$

Hence, we can choose a positive constant $K_0$ such that

$$\left[\Phi_n\left(\sqrt{m}\mu(d_{m,n} + \eta m^{-1/(2k)})/(\sqrt{2}\sigma_0)\right) - \gamma\right] > K_0$$

for all sufficiently large $n$ and thus, from (B.5) we get

$$M_{m,n}(d) - M_{m,n}(d_{m,n}) \geq f_0 K_0 \left(d - (d_{m,n} + \eta m^{-1/(2k)})\right) \tag{B.6}$$

for $d \in (d_{m,n} + \eta m^{-1/(2k)}, d_0 + \epsilon)$. Also, $|\Phi_n(0) - \gamma| > (1/2)\,|1/2 - \gamma|$, for large $n$. Choose $K_1 = \frac{1}{2} f_0 \; \min\,[K_0, |\gamma - 1/2|]$ in (B.1). Then,

$$\begin{aligned}
[\rho_n(d, d_{m,n}) < \kappa] &= [d_0, d_{m,n} + \eta m^{-1/(2k)}] \cup [d < d_0, |d - d_0| < \kappa^2/K_1] \\
&\quad \cup [d > d_{m,n} + \eta m^{-1/(2k)}, |d - d_{m,n} - \eta m^{-1/(2k)}| < \kappa^2/K_1] \\
&\subset [|d - d_{m,n}| < \kappa^2/K_1 + A m^{-1/(2k)}].
\end{aligned}$$

Here, $A$ is a fixed constant chosen such that $A > \max(\eta, m^{-1/(2k)}(d_{m,n} - d_0))$, for all sufficiently large $n$; this follows from (3.7). Let $\kappa$ be chosen such that $\kappa^2/K_1 + 2Am^{-1/(2k)} < \epsilon$ for all sufficiently large $n$. As $|d_0 - d_{m,n}| < Am^{-1/(2k)}$, this gives $[\rho_n(d, d_{m,n}) < \kappa] \subset (d_0 - \epsilon, d_0 + \epsilon)$. Thus, for large $n$ and $d$ such that $[\rho_n(d, d_{m,n}) < \kappa]$, using the definition of $\rho_n$ and relations (B.4) and (B.6), we have the desired result.

## B.2 Proof of Theorem 3.6

In order to deduce the limit of the process $\hat{V}_n$ (see (3.8)), we first prove a lemma that allows us to work with $\sigma_0$ instead of $\hat{\sigma}$.

**Lemma B.3.** *Let* $V_n(t) = n\{\mathbb{M}_{m,n}(d_0 + t/n, \sigma_0) - \mathbb{M}_{m,n}(d_0, \sigma_0)\}$. *Then, for any* $L > 0$,

$$\sup_{t \in [-L, L]} |\hat{V}_n(t) - V_n(t)| \xrightarrow{P} 0,$$

*where* $\xrightarrow{P}$ *denotes convergence in probability.*

*Proof.* For all $t \in [-L, L]$, we have

$$
\begin{aligned}
&|\hat{V}_n(t) - V_n(t)| \\
= \;& \left| \sum_{i=1}^{n} \left\{ \Phi\left(\frac{\sqrt{m}\bar{Y}_i}{\hat{\sigma}}\right) - \Phi\left(\frac{\sqrt{m}\bar{Y}_i}{\sigma_0}\right) \right\} \left( 1\left(X_i \le d_0 + \frac{t}{n}\right) - 1\left(X_i \le d_0\right) \right) \right| \\
\le \;& \sup_{y \in \mathbb{R}} \left| \Phi\left(\frac{\sqrt{m}y}{\hat{\sigma}}\right) - \Phi\left(\frac{\sqrt{m}y}{\sigma_0}\right) \right| \sum_{i=1}^{n} 1\left( X_i \in \left[ d_0 - \frac{L}{n}, d_0 + \frac{L}{n} \right] \right) \\
\le \;& \sup_{u \in \mathbb{R}} \left| \Phi\left(u\right) - \Phi\left(\frac{\hat{\sigma}}{\sigma_0} u\right) \right| \sum_{i=1}^{n} 1\left( X_i \in \left[ d_0 - \frac{L}{n}, d_0 + \frac{L}{n} \right] \right).
\end{aligned}
$$

Also, $\sigma \mapsto \sup_{u \in \mathbb{R}} |\Phi(u) - \Phi(u\sigma/\sigma_0)|$ can be shown to be continuous; in fact, a closed form expression can be obtained by taking derivatives. It can be seen that for $a \in (0, \infty)$,

$$
\sup_{u \in \mathbb{R}} |\Phi(u) - \Phi(au)| = \begin{cases} 0, & a = 1, \\ \left| \Phi\left(\sqrt{\frac{2 \ \log a}{a^2 - 1}}\right) - \Phi\left(a\sqrt{\frac{2 \ \log a}{a^2 - 1}}\right) \right|, & a \neq 1. \end{cases}
$$

This can be shown to be continuous at 1 by elementary calculations. Thus the first term in the bound for $|\hat{V}_n(t) - V_n(t)|$ converges in probability to 0. Moreover, the remaining term is a Binomial random variable $(Bin(n, F(d_0 + L/n) - F(d_0 - L/n)))$ which converges weakly to the Poisson distribution with parameter $2Lf(d_0)$. Thus by Slutsky's theorem, we obtain the desired result. □

We now continue with the proof of Theorem 3.6. We first prove that $(\hat{V}_n, J_n)$ converges weakly to $(V, J)$ as processes in $D[-C, C] \times D[-C, C]$, for each positive integer $C$. By Lemma B.3, it suffices to show that $(V_n, J_n)$ converges weakly to $(V, J)$.

To justify the finite dimensional convergence of $(V_n, J_n)$ to $(V, J)$, first on $[0, \infty)$, let $0 = t_0 \le t_1 < t_2 < \ldots < t_l$. By Cramér-Wold device, it suffices to show that the characteristic function of

$$(V_n(t_1), J_n(t_1), V_n(t_2) - V_n(t_1), J_n(t_2) - J_n(t_1), \ldots, V_n(t_l) - V_n(t_{l-1}), J_n(t_l) - J_n(t_{l-1}))$$

converges to that of

$$(V(t_1), J(t_1), V(t_2) - V(t_1), J(t_2) - J(t_1), \ldots, V(t_l) - V(t_{l-1}), J(t_l) - J(t_{l-1}))\,.$$

We illustrate this derivation for $l = 2$, the extension to larger $l$s following in a straightforward manner. For $(c_i, d_i) \in \mathbb{R}^2$, $i = 1, 2$, consider the expression

$$E\left[\exp\left[\imath\left(c_1 V_n(t_1) + d_1 J_n(t_1) + \{c_2(V_n(t_2) - V_n(t_1)) + d_2(J_n(t_2) - J_n(t_1))\}\right)\right]\right]. \quad \text{(B.7)}$$

As $t_0 = 0$, note that

$$\begin{aligned} & c_1 V_n(t_1) + d_1 J_n(t_1) + c_2(V_n(t_2) - V_n(t_1)) + d_2(J_n(t_2) - J_n(t_1)) \\ = & \sum_{j=1}^{n}\sum_{i=1}^{2}\left\{c_i \Phi\left(\frac{\sqrt{m}\bar{Y}_j}{\sigma_0}\right) - c_i\gamma + d_i\right\} 1\left(X_j \in \left(d_0 + \frac{t_{i-1}}{n}, d_0 + \frac{t_i}{n}\right]\right). \end{aligned}$$

The above summands are independent for different $j$s and hence, (B.7) equals

$$\left[E\left[\exp\left[\imath\sum_{i=1}^{2}\left\{c_i \Phi\left(\frac{\sqrt{m}\bar{Y}_1}{\sigma_0}\right) - c_i\gamma + d_i\right\} 1\left(X_1 \in \left(d_0 + \frac{t_{i-1}}{n}, d_0 + \frac{t_i}{n}\right]\right)\right]\right]\right]^n.$$

Let $Z_{1n}$ be as defined in (3.4) and $Z \sim N(0,1)$. Taking iterated expectations (by first conditioning on $X_1$), the above display equals $(1 + \xi_n/n)^n$, where

$$\begin{aligned} \xi_n &= n\sum_{i=1}^{2}\int_{d_0+t_{i-1}/n}^{d_0+t_i/n}\left[E\left[\exp\left(\imath\left\{c_i\Phi\left(\frac{\sqrt{m}\mu(x)}{\sigma_0} + Z_{1n}\right) - c_i\gamma + d_i\right\}\right)\right] - 1\right] f(x)dx \\ &= \sum_{i=1}^{2}\int_{t_{i-1}}^{t_i}\left[\left[E\left(c_i \exp\left(\imath s\left\{\Phi\left(\frac{\sqrt{m_0}\;\mu^{(k)}(d_0+)}{k!\sigma_0}u^k + o(1) + Z_{1n}\right) - c_i\gamma + d_i\right\}\right)\right) - 1\right]\right. \end{aligned}$$

$$\times f\left(d_0+\frac{u}{n}\right)\Big]\, du.$$

The $o(1)$ term appearing in the above expression does not depend on $u$ as

$$\sup_{(d_0, d_0+\zeta_0)} |\mu^{(k)}(x)| < \infty$$

by Assumption 1. As $Z_{1n} + o(1)$ converges weakly to $Z$ and $\exp(\imath \cdot)$ is bounded, $\xi_n$ converges to $f(d_0)\xi_0$ where

$$\xi_0 = \sum_{i=1}^{2} \int_{t_{i-1}}^{t_i} \left[E\left(\exp\left(\imath s\left\{\Phi\left(\frac{\sqrt{m_0}\ \mu^{(k)}(d_0+)}{k!\sigma_0}u^k + Z\right) - \gamma\right\}\right)\right) - 1\right] du.$$

So, the expression in (B.7) converges to $\exp\left(f(d_0)\xi_0\right)$. This is precisely the characteristic function of $(V(t_1), J(t_1), V(t_2) - V(t_1), J(t_2) - J(t_1))$ evaluated at $(c_1, d_1, c_2, d_2)$. To see this, first note that $(V(t_1), J(t_1))$ and $(V(t_2) - V(t_1), J(t_2) - J(t_1))$ are independent by virtue of the fact that the arrival times of events occurring over disjoint sets are independent for a Poisson process. Further, let $W_j$ be i.i.d. $U(0,t)$, for $j \geq 1$, which are independent of $\{Z_j\}_{j\geq 1}$ and $\nu^+$. Using the order statistic characterization of the arrival times of a Poisson process,

$$\begin{aligned}
& E\left[\exp\left(\imath\left\{c_1 V(t_1) + d_1 J(t_1)\right\}\right) | \nu^+(t_1)\right] \qquad \text{(B.8)}\\
= \;& E\left[\exp\left(\sum_{j=1}^{\nu^+(t)} \imath\left(c_1\Phi\left(\frac{\sqrt{m_0}\ \mu^{(k)}(d_0+)}{k!\sigma_0}S_j^k + Z_j\right) - c_1\gamma + d_1\right)\right)\middle|\, \nu^+(t_1)\right]\\
= \;& E\left[\exp\left(\sum_{j=1}^{\nu^+(t)} \imath\left(c_1\Phi\left(\frac{\sqrt{m_0}\ \mu^{(k)}(d_0+)}{k!\sigma_0}W_j^k + Z_j\right) - c_1\gamma + d_1\right)\right)\middle|\, \nu^+(t_1)\right]\\
= \;& \left[E\ \exp\left(\imath\left(c_1\Phi\left(\frac{\sqrt{m_0}\ \mu^{(k)}(d_0+)}{k!\sigma_0}W_1^k + Z_1\right) - c_1\gamma + d_1\right)\right)\right]^{\nu^+(t_1))}\\
= \;& \left[\frac{g(c_1, d_1, 0, t_1) + t_1}{t_1}\right]^{\nu^+(t_1)},
\end{aligned}$$

where for $0 \leq s < t$,

$$g(c_1, d_1, s, t) = \int_s^t \left[ E\left( \exp\left( \imath \left\{ c_1 \Phi\left( \frac{\sqrt{m_0}\, \mu^{(k)}(d_0+)}{k!\sigma_0} u^k + Z \right) - c_1\gamma + d_1 \right\} \right) \right) - 1 \right] du.$$

Note that the relation in (B.8) holds even when $\nu^+(t)$ is 0. Thus,

$$E\left[\exp\left(\imath \left\{c_1 V(t_1) + d_1 J(t_1)\right\}\right)\right] = \exp(f(d_0) g(c_1, d_1, 0, t_1)).$$

Similarly, it can be deduced that

$$E\left[\exp\left(\imath \left\{c_2 (V(t_2) - V(t_1)) + d_2 (J(t_2) - J(t_1))\right\}\right)\right] = \exp(f(d_0) g(c_2, d_2, t_1, t_2)).$$

Using the independence between $(V(t_1), J(t_1))$ and $(V(t_2) - V(t_1), J(t_2) - J(t_1))$, we get that the limit of (B.7) is indeed the characteristic function of

$$\left(V(t_1), J(t_1), V(t_2) - V(t_1), J(t_2) - J(t_1)\right).$$

Hence, finite dimensional convergence of $(V_n, J_n)$ to $(V, J)$ on $[0, \infty]$ follows from Lévy continuity theorem. The finite dimensional convergence on the entire domain can be deduced analogously.

Next, we complete the proof of weak convergence of $(V_n, J_n)$ to $(V, J)$ by showing asymptotic tightness. For $t_1 < t < t_2$ and sufficiently large $n$,

$$\begin{aligned}
& E[|J_n(t) - J_n(t_1)||J_n(t_2) - J_n(t)|] \\
= \;& E\left[ \sum_{i=1}^{n} \sum_{j=1}^{n} 1\left( X_i \in \left( d_0 + \frac{t_1}{n}, d_0 + \frac{t}{n} \right] \right) 1\left( X_j \in \left( d_0 + \frac{t}{n}, d_0 + \frac{t_2}{n} \right] \right) \right] \\
= \;& n(n-1) E\left[ \left( X_1 \in \left( d_0 + \frac{t_1}{n}, d_0 + \frac{t}{n} \right] \right) 1\left( X_2 \in \left( d_0 + \frac{t}{n}, d_0 + \frac{t_2}{n} \right] \right) \right] \\
\leq \;& 2\left\|f\right\|_\infty^2 (t - t_1)(t_2 - t) \leq 2\left\|f\right\|_\infty^2 (t_2 - t_1)^2,
\end{aligned}$$

where $\|f\|_\infty < \infty$ by Assumption 2. The above relation shows that the condition stated for tightness in Theorem 15.6 of Billingsley (1968, pp. 128) is satisfied and hence, the process $J_n$ is asymptotically tight. As $|V_n(t) - V_n(t_1)| \leq |J_n(t) - J_n(t_1)|$, the process $V_n$ is asymptotically tight. As both the marginal processes are tight, $(V_n, J_n)$ is tight and hence, condition (i) of Theorem 3.5 is satisfied.

Moreover, no two flat stretches of $V(t), t \in [-C, C]$, have the same height (w.p. 1). To see this, let $A_C$ denote this event and define $R_i = \sum_{j=1}^{i} \left(\Phi\left(\sqrt{m_0}\, \mu^{(k)}(d_0+)/(k!\sigma_0) S_j^k + Z_j\right) - \gamma\right)$ when $i > 0$, and $R_i = \sum_{j=1}^{-i} (\gamma - U_j)$ when $i < 0$, and $R_0 = 0$. For non-negative integers $n_1$ and $n_2$, $n_1 + n_2 > 0$, we have $P[R_i = R_l | \nu^+(C) = n_1, \nu^-(C) = n_2] = 0$ for $n_1 \geq i > l \geq -n_2$. This is because given $\nu^+(C) = n_1$ and $\nu^-(C) = n_2$, the arrival times for $S_j$s are the order statistics from $U(0, C)$ and thus $R_i - R_l$ is a continuous random variable. Now,

$$\begin{aligned} & P[A_C | \nu^+(C) = n_1, \nu^-(C) = n_2] \\ = & \; 1 - P\left[\bigcup_{n_1 \geq i > l \geq -n_2} [R_i = R_l] \middle| \nu^+(C) = n_1, \nu^-(C) = n_2\right] = 1. \end{aligned}$$

Also, $P[A_C | \nu^+(C) = 0, \nu^-(C) = 0] = 1$. Hence,

$$P[A_C] = E[P[A_C | \nu^+(C), \nu^-(C)]] = 1.$$

Further, let $\hat{h}_l = n(\hat{d}_{m,n} - d_0)$ and $\hat{h}_u$ denote the smallest and largest minimizers of $\hat{V}_n(t)$, respectively. Using Theorem 3.2, $(\hat{h}_l, \hat{h}_u)$ is $O_P(1)$. Also, let $h_l$ and $h_u$ denote the smallest and largest minimizers for $V(t)$. As $V(0) = 0$ and $V(t) \to \infty$ as $|t| \to \infty$ w.p. 1, we get $(h_l, h_u) = O_P(1)$. To see that $V(t) \to \infty$ as $|t| \to \infty$ a.s., note that $\sum_{j=1}^{n} (\gamma - U_j)/n \to \gamma - \frac{1}{2} > 0$ and $\nu^-(-t) \to \infty$, a.s. So, we get $V(t) \to \infty$ as $t \to \infty$ a.s. Also, choose $\epsilon > 0$ and $\eta_\epsilon > 0$ such that $\gamma + \epsilon < 1$ and $E\Phi[\eta_\epsilon + Z_1] = \Phi[\eta_\epsilon/\sqrt{2}] = \gamma + \epsilon$. Then by the SLLN, $\sum_{j=1}^{n} (\Phi[\eta_\epsilon + Z_j] - \gamma)/n \to \epsilon$ a.s. As $S_j \to \infty$ and $\nu^+(t) \to \infty$ a.s., we get $\liminf_{t \to \infty} \{V(t)/\nu^+(t)\} \geq \epsilon$ a.s. Thus

$V(t) \to \infty$ as $|t| \to \infty$ w.p. 1. Hence, by applying Theorem 3.5 we get the desired result. $\square$

## B.3 Proof of Proposition 3.8

The proof of Proposition 3.8 follows along the same lines as that of Theorem 3.6. Here, we briefly justify the form of the limiting distribution. By calculations analogous to those used for simplifying (B.7), it can be shown that for $t > 0$, $E(\exp(\imath c V_n(t))) = \left(1 + \bar{\xi}_n/n\right)^n$, where

$$\begin{aligned} \bar{\xi}_n &= \int_0^t \left[\left\{E\left(\exp\left(\imath c\left\{\Phi\left(\frac{\sqrt{m}\mu(d_0+u/n)}{\sigma_0} + Z_{1n}\right) - \gamma\right\}\right)\right) - 1\right\}\right. \\ &\quad \left. \times f\left(d_0 + \frac{u}{n}\right)\right] du \\ &\to f(d_0)\int_0^t \left[\exp\left(\imath c\left\{1-\gamma\right\}\right) - 1\right] du = f(d_0)\{\exp(\imath c(1-\gamma)) - 1\}t. \end{aligned}$$

The above convergence uses the fact $\sqrt{m}\mu(d_0 + u/n)/\sigma_0 \to \infty$ for $u > 0$, which can be justified through a $k$-th order Taylor expansion of $\mu$ around $d_0$. The limit here is precisely the characteristic function of $\bar{V}(t)$. Hence, the one-dimensional marginals of $V_n$ converge to that of $\bar{V}$ on the positive half line. The remainder of the proof is almost identical to that for Theorem 3.6.

## B.4 Proof of Proposition 3.9

For proving Proposition 3.9, we first prove the following lemma to justify imputing $\sigma_0$ in place of $\hat{\sigma}$ in the local processes.

**Lemma B.4.** *Consider the case when $\beta < 2k$. Let*

$$H_n(t) = n^{\beta/(2k)}\left\{\mathbb{M}_{m,n}\left(d_0 + t/n^{\beta/(2k)}, \sigma_0\right) - \mathbb{M}_{m,n}(d_0, \sigma_0)\right\}.$$

*Then, for any $L > 0$,*

$$\sup_{t \in [-L,L]} |\hat{H}_n(t) - H_n(t)| \xrightarrow{P} 0.$$

*Proof.* For $t \in [-L, L]$,

$$\begin{aligned}
&|\hat{H}_n(t) - H_n(t)| \\
&= n^{\beta/(2k)-1} \left| \sum_{i=1}^{n} \left\{ \Phi\left(\frac{\sqrt{m}\bar{Y}_i}{\hat{\sigma}}\right) - \Phi\left(\frac{\sqrt{m}\bar{Y}_i}{\sigma_0}\right) \right\} \times \right. \\
&\quad \left. \left( 1\left(X_i \le d_0 + \frac{t}{n^{\beta/(2k)}}\right) - 1\left(X_i \le d_0\right) \right) \right| \\
&\le n^{\beta/(2k)-1} \sup_{y \in \mathbb{R}} \left| \Phi\left(\frac{\sqrt{m}y}{\hat{\sigma}}\right) - \Phi\left(\frac{\sqrt{m}y}{\sigma_0}\right) \right| \times \\
&\quad \sum_{i=1}^{n} 1\left( X_i \in \left[ d_0 - \frac{L}{n^{\beta/(2k)}}, d_0 + \frac{L}{n^{\beta/(2k)}} \right] \right) \\
&\le \sup_{u \in \mathbb{R}} \left| \Phi(u) - \Phi\left(\frac{\hat{\sigma}}{\sigma_0} u\right) \right| \left\{ n^{\beta/(2k)-1} \sum_{i=1}^{n} 1\left( X_i \in \left[ d_0 - \frac{L}{n^{\beta/(2k)}}, d_0 + \frac{L}{n^{\beta/(2k)}} \right] \right) \right\}.
\end{aligned}$$

As in the proof of Lemma B.3, the first term goes in probability to zero. As for the second term,

$$\begin{aligned}
&\mathrm{Var}\left[ n^{\beta/(2k)-1} \sum_{i=1}^{n} 1\left( X_i \in \left[ d_0 - \frac{L}{n^{\beta/(2k)}}, d_0 + \frac{L}{n^{\beta/(2k)}} \right] \right) \right] \\
&\quad = n^{2(\beta/(2k)-1)} n \ \mathrm{Var}\left[ 1\left( X_i \in \left[ d_0 - \frac{L}{n^{\beta/(2k)}}, d_0 + \frac{L}{n^{\beta/(2k)}} \right] \right) \right] \\
&\quad = n^{2(\beta/(2k)-1)} n \ O(n^{-\frac{\beta}{k}}) = O(n^{-1}) \to 0,
\end{aligned}$$

and

$$\begin{aligned}
&E\left[ n^{\beta/(2k)-1} \sum_{i=1}^{n} 1\left( X_i \in \left[ d_0 - \frac{L}{n^{\beta/(2k)}}, d_0 + \frac{L}{n^{\beta/(2k)}} \right] \right) \right] \\
&\quad = n^{(\beta/(2k)-1)} n \ E\left[ 1\left( X_i \in \left[ d_0 - \frac{L}{n^{\beta/(2k)}}, d_0 + \frac{L}{n^{\beta/(2k)}} \right] \right) \right] \\
&\quad = n^{(\beta/(2k)-1)} n \ O(n^{-\beta/(2k)}) = O(1).
\end{aligned}$$

Thus the second term is $O(1)+o_P(1)$. Hence, we get the result. $\square$

We use a version of the Arzela-Ascoli theorem in several proofs and thus we state it below for convenience.

**Theorem B.5** (Arzela-Ascoli). *Let $f_n$ be a sequence of continuous functions defined on a compact set $[a,b]$ such that $f_n$ converge pointwise to $f$ and for any $\delta_n \downarrow 0$ $\sup_{|x-y|<\delta_n} |f_n(x)-f_n(y)|$ converges to 0. Then $\sup_{x\in[a,b]} |f_n(x)-f(x)|$ converges to zero.*

We now continue with the proof of Proposition 3.9. Using Lemma B.4, for proving (3.10), it would suffice to show that

$$\sup_{t\in[-L,L]} |H_n(t)-c(t)| \xrightarrow{P} 0. \tag{B.9}$$

Let

$$c_n(t) = E\{H_n(t)\} = n^{\beta/(2k)} \int_{d_0}^{d_0+tn^{-\beta/(2k)}} E\left[\left\{\Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma_0}\right)-\gamma\right\}\middle| X=x\right] f(x)dx.$$

For $x<0$ and given $X=x$, $\Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma_0}\right) \xrightarrow{d} U(0,1)$. Hence, by the dominated convergence theorem (DCT), $c_n(t) \to \left(\frac{1}{2}-\gamma\right) f(d_0)t$, for $t \le 0$. For $t>0$, we have:

$$\begin{aligned}
c_n(t) &= n^{\beta/(2k)} \int_{d_0}^{d_0+tn^{-\beta/(2k)}} E\left[\left\{\Phi\left(\frac{\sqrt{m}\bar{Y}}{\sigma_0}\right)-\gamma\right\}\middle| X=x\right] f(x)dx \\
&= n^{\beta/(2k)} \int_{d_0}^{d_0+tn^{-\beta/(2k)}} \left\{\Phi\left(\frac{\sqrt{m}\mu(x)}{\sqrt{2}\sigma_0}\right)-\gamma\right\} f(x)dx \\
&= \int_0^t \left\{\Phi\left(\frac{\sqrt{m}\mu(d_0+u/n^{\beta/(2k)})}{\sqrt{2}\sigma_0}\right)-\gamma\right\} f\left(d_0+\frac{u}{n^{\beta/(2k)}}\right) du
\end{aligned}$$

$$
\begin{aligned}
&= \int_0^t \left\{ \Phi\left( \frac{\sqrt{m_0}\ \mu^{(k)}(d_0+)}{\sqrt{2}k!\sigma_0} u^k + o(1) \right) - \gamma \right\} f\left(d_0 + \frac{u}{n^{\beta/(2k)}}\right) du \\
&\to f(d_0) \int_0^t \left\{ \Phi\left( \frac{\sqrt{m_0}\ \mu^{(k)}(d_0+)}{\sqrt{2}k!\sigma_0} u^k \right) - \gamma \right\} du, \text{ by DCT.}
\end{aligned}
$$

Hence, $c_n(t) \to c(t)$. In fact, this convergence is uniform on any compact set. To see this, note that $|c_n(t) - c_n(s)| \le \|f\|_\infty |t-s|$. So, $c_n$s are equicontinuous and thus by Arzela-Ascoli, the convergence is uniform on $[-L, L]$ for every $L > 0$. Further, let $\tilde{H}_n(t) = n^{1/2-\beta/(4k)}(H_n(t) - c_n(t))$. Then, for $t_1 < t < t_2$,

$$
\begin{aligned}
E|\tilde{H}_n(t) - \tilde{H}_n(t_1)|^2 |\tilde{H}_n(t_2) - \tilde{H}_n(t)|^2 &= E\ |\tilde{H}_n(t) - \tilde{H}_n(t_1)|^2\ E\ |\tilde{H}_n(t_2) - \tilde{H}_n(t)|^2 \\
= \ & \mathrm{Var}\left[ n^{\beta/(4k)} \left\{ \Phi\left( \frac{\sqrt{m}\bar{Y}}{\sigma_0} \right) - \gamma \right\} 1\left( X_1 \in \left( d_0 + \frac{t_1}{n^{\beta/(2k)}}, d_0 + \frac{t}{n^{\beta/(2k)}} \right] \right) \right] \times \\
& \mathrm{Var}\left[ n^{\beta/(4k)} \left\{ \Phi\left( \frac{\sqrt{m}\bar{Y}}{\sigma_0} \right) - \gamma \right\} 1\left( X_1 \in \left( d_0 + \frac{t}{n^{\beta/(2k)}}, d_0 + \frac{t_2}{n^{\beta/(2k)}} \right] \right) \right] \\
\le \ & n^{\beta/(2k)} E\left[ \left\{ \Phi\left( \frac{\sqrt{m}\bar{Y}}{\sigma_0} \right) - \gamma \right\} 1\left( X_1 \in \left( d_0 + \frac{t_1}{n^{\beta/(2k)}}, d_0 + \frac{t}{n^{\beta/(2k)}} \right] \right) \right]^2 \times \\
& n^{\beta/(2k)} E\left[ \left\{ \Phi\left( \frac{\sqrt{m}\bar{Y}}{\sigma_0} \right) - \gamma \right\} 1\left( X_1 \in \left( d_0 + \frac{t}{n^{\beta/(2k)}}, d_0 + \frac{t_2}{n^{\beta/(2k)}} \right] \right) \right]^2 \\
\le \ & \|f\|_\infty^2 (t - t_1)(t_2 - t) \le \|f\|_\infty^2 (t_2 - t_1)^2.
\end{aligned}
$$

So, by Theorem 15.6 of Billingsley (1968, pp. 128), $\tilde{H}$ is tight in $D(\mathbb{R})$. As $\beta < 2k$, $(H_n(t) - c_n(t)) \xrightarrow{d} 0$ and hence $H_n(t) \xrightarrow{d} c(t)$ as processes in $D(\mathbb{R})$. As the limiting process in degenerate and $x(\cdot) \mapsto \sup_{t \in [-L,L]} |x(t)|$ is continuous, we get (B.9).

Moreover the limit process, $c(t)$, is continuous and has a unique minimum. Also, $n^{\beta/(2k)}(\hat{d}_{m,n} - d_0)$ is $O_P(1)$. Thus, by the argmin continuous mapping, we obtain the desired result. $\square$

## B.5 Proof of Proposition 3.11

We first show that

$$\frac{1}{nh_n^{2k+1}} g_1(\hat{d}_{m,n}, h_n) = \frac{f(d_0)}{2k+1} + o_P(1),$$

where

$$g_1(d,h) = \sum_{i=1}^{n} (X_i - d)^{2k} 1(X_i \in (d, d+h]).$$

Note that $h_n^{-(2k+1)}(\hat{d}_{m,n} - d_0) = o_P(1)$. Fix $\delta > 0$. Then $P[|\hat{d}_{m,n} - d_0| < \delta\ h_n^{2k+1}]$ converges to 1. On the set $[|\hat{d}_{m,n} - d_0| < \delta\ h_n^{2k+1}]$,

$$g_1(d_0 - \delta h_n^{2k+1}, h_n + 2\delta h_n^{2k+1}) \geq g_1(\hat{d}_{m,n}, h_n) \geq g_1(d_0 + \delta h_n^{2k+1}, h_n - 2\delta h_n^{2k+1}). \quad \text{(B.10)}$$

So, it suffices to show that the above two bounds converge in probability to $f(d_0)/(2k+1)$. Note that

$$\begin{aligned}
& E\left[\frac{1}{nh_n^{2k+1}} g_1(d_0 + \delta h_n^{2k+1}, h_n - 2\delta h_n^{2k+1})\right] \\
&= \frac{n}{nh_n^{2k+1}} \int_{d_0+\delta h_n^{2k+1}}^{d_0+h_n-\delta h_n^{2k+1}} (x - d_0 - \delta h_n^{2k+1})^{2k} f(x) dx \\
&= \frac{1}{h_n^{2k+1}} \int_0^{1-\delta h_n^{2k}} (uh_n)^{2k} f(d_0 + \delta h_n^{2k+1} + uh_n) h_n du \\
&= f(d_0) \int_0^1 u^{2k} du + o(1) = \frac{f(d_0)}{2k+1} + o(1),
\end{aligned}$$

and

$$\begin{aligned}
& \mathrm{Var}\left[\frac{1}{nh_n^{2k+1}} g_1(d_0 + \delta h_n^{2k+1}, h_n - 2\delta h_n^{2k+1})\right] \\
&= \frac{n}{(nh_n^{2k+1})^2} \mathrm{Var}\left[(X_1 - d_0 - \delta h_n^{2k+1})^{2k} 1(X_1 \in (d_0 + \delta h_n^{2k+1}, d_0 - \delta h_n^{2k+1} + h_n])\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{n}{(nh_n^{2k+1})^2} E\left[(X_1 - d_0 - \delta h_n^{2k+1})^{4k} 1(X_1 \in (d_0 + \delta h_n^{2k+1}, d_0 - \delta h_n^{2k+1} + h_n])\right] \\
&\leq \frac{n}{(nh_n^{2k+1})^2}(h_n - 2\delta h_n^{2k+1})^{4k+1}(f(d_0) + o(1)) = \frac{O(1)}{nh_n} \to 0.
\end{aligned}
$$

Thus,

$$
\frac{1}{nh_n^{2k+1}} g_1(d_0 + \delta h_n^{2k+1}, h_n - 2\delta h_n^{2k+1}) = \frac{f(d_0)}{2k+1} + o_P(1).
$$

The treatment of the upper bound in (B.10) is similar. Next, let $g_2(d,h) = \sum \bar{Y}_i(X_i - d)^k 1(X_i \in (d, d+h])$. As the $k$-th derivative of $\mu$ is bounded in $(d_0, d_0 + \zeta)$ for sufficiently small $\zeta$, we have

$$
\begin{aligned}
E\left[\frac{1}{nh_n^{2k+1}} g_2(d_0, h_n)\right] &= \frac{n}{nh_n^{2k+1}} \int_{d_0}^{d_0+h_n} \mu(x)(x-d_0)^k f(x)dx \\
&= \frac{1}{h_n^{2k+1}} \int_0^1 \mu(d_0 + uh_n)(uh_n)^k f(d_0 + \delta h_n^{2k+1} + uh_n) h_n du \\
&= \frac{1}{h_n^{2k+1}} \int_0^1 (\xi(uh_n)^k + o((uh_n)^k))(uh_n)^k f(d_0 + \delta h_n^{2k+1} + uh_n) h_n du \\
&= \xi \frac{f(d_0)}{2k+1} + o(1), \text{by DCT.}
\end{aligned}
$$

Also, by similar calculations,

$$
\begin{aligned}
&\mathrm{Var}\left[\frac{1}{nh_n^{2k+1}} g_2(d_0, h_n)\right] \\
&\leq \frac{n}{(nh_n^{2k+1})^2} E\left[\bar{Y}_1(X_1 - d_0)^k 1(X_1 \in (d_0, d_0 + h_n])\right]^2 \\
&= \frac{n}{(nh_n^{2k+1})^2} E\left[\left\{\mu^2(X_1) + \sigma_0^2/(m_0 n^{2k})\right\}(X_1 - d_0)^{2k} 1(X_1 \in (d_0, d_0 + h_n])\right] \\
&\leq \frac{O(1)}{nh_n} + \frac{n\sigma_0^2}{c(nh_n^{2k+1})^2 n^{2k}} O(h_n^{2k+1}) \to 0.
\end{aligned}
$$

So, $(1/(nh_n^{2k+1}))g_2(d_0,h_n) = \xi(f(d_0)/(2k+1)) + o_P(1)$. To conclude the final result, we need to show that

$$\frac{1}{nh_n^{2k+1}}\left\{g_2(\hat{d}_{m,n},h_n) - g_2(d_0,h_n)\right\} = o_P(1).$$

Let $M_0 = \sup_{d\in(d_0,d_0+\zeta)}\mu(d)$, which is finite for sufficiently small $\zeta$. On the set $[|\hat{d}_{m,n} - d_0| < \delta h_n^{2k+1}]$, and for large $n$,

$$\begin{aligned}
&\left|g_2(\hat{d}_{m,n},h_n) - g_2(d_0,h_n)\right| \\
\leq\ & \sup_{|d-d_0|<\delta h_n^{2k+1}}\left|\frac{1}{nh_n^{2k+1}}\{g_2(d,h_n) - g_2(d_0,h_n)\}\right| \\
\leq\ & \sup_{|d-d_0|<\delta h_n^{2k+1}}\sum_{i=1}^{n}\Big[|\bar{Y}_i|(X_i-d)^k - (X_i-d_0)^k|1(X_i\in(d_0,d_0+h_n]\cap(d,d+h_n]) \\
&\quad + \ |\bar{Y}_i|(X_i - d\wedge d_0)^k 1(X_i\in(d_0,d_0+h_n]\Delta(d,d+h_n])\Big] \\
\leq\ & \sup_{|d-d_0|<\delta h_n^{2k+1}}\sum_{i=1}^{n}\Big[|\bar{Y}_i|k(X_i-d_0+\delta h_n^{2k+1})^{k-1}|d-d_0|1(X_i\in(d_0,d_0+h_n]) \\
&\quad + |\bar{Y}_i|(X_i-d_0-\delta h_n^{2k+1})^k\Big\{1(|X_i-d_0|\leq\delta h_n^{2k+1}) + \\
&\quad 1(|X_i-d_0-h_n|\leq\delta h_n^{2k+1})\Big\}\Big] \\
=\ & O(h_n^{3k})\sum_{i=1}^{n}|\bar{Y}_i|1(X_i\in(d_0,d_0+h_n]) \\
&\quad + O(h_n^k)\sum_{i=1}^{n}|\bar{Y}_i|\{1(|X_i-d_0|\leq\delta h_n^{2k+1}) + 1(|X_i-d_0-h_n|\leq\delta h_n^{2k+1})\} \\
\leq\ & O(h_n^{3k})\sum_{i=1}^{n}(M_0+|\bar{\epsilon}_i|)1(X_i\in(d_0,d_0+h_n]) \\
&\quad + O(h_n^k)\sum_{i=1}^{n}(M_0+|\bar{\epsilon}_i|)\{1(|X_i-d_0|\leq\delta h_n^{2k+1}) + 1(|X_i-d_0-h_n|\leq\delta h_n^{2k+1})\} \\
\leq\ & O(h_n^{3k})O_P(nh_n) + O(h_n^k)O_P(nh_n^{2k+1}) = o_P(nh_n^{2k+1}).
\end{aligned}$$

The last inequality follows from the fact that $(1/(nh_n))\sum_{i=1}^{n}(M_0+|\bar{\epsilon}_i|)1(X_i \in (d_0,d_0+h_n])$ converges in probability to $M_0 f(d_0)$, which can be justified by computing the limiting means and variances. This completes the proof. $\square$

## B.6 Proof of Proposition 3.14

For $\epsilon > 0$ and $x \in \mathbb{R}$, let

$$U_n(x) = U_n(x, \beta) = \frac{1}{N_n} \sum_{j=1}^{N_n} 1 \left[ q_n(\hat{d}_{n,q_n,j} - d_0) \leq x \right]$$

and $E_n = [q_n|\hat{d}_n - d_0| \leq \epsilon]$. As $q_n/n \to 0$ and $n(\hat{d}_n - d_0) = O_P(1)$, $P(E_n) \to 1$. Moreover, on the set $E_n$,

$$U_n(x - \epsilon) \leq L_{n,q}(x) \leq U_n(x + \epsilon).$$

Hence, to show pointwise convergence (in probability) of $L_{n,q}(\cdot, \beta)$ to $G_\beta(\cdot)$, it suffices to show that $U_n(x, \beta) \xrightarrow{P} G_\beta(x)$. Note that $E\left[U_n(x)\right] = G_{n,\beta}(x) \to G_\beta(x)$. So, it suffices to show that $\mathrm{Var}(U_n(x)) \to 0$. To this end, let $s_n = \lfloor n/q_n \rfloor$. For $j = 0, \ldots, (s_n - 1)$, let $R_{n,q_n,j}$ be the statistic $\hat{d}_{q_n}$ computed from the data set $(X_{q_n j+1}, Y_{(q_n j+1)\ 1}, \ldots, Y_{(q_n j+1)\ l_n}; \ldots; X_{q_n j+q_n}, Y_{(q_n j+q_n)\ 1}, \ldots, Y_{(q_n j+q_n)\ l_n})$ and

$$\bar{U}_n(x) = \frac{1}{s_n} \sum_{j=1}^{s_n} 1 \left[ q_n(R_{n,q_n,j} - d_0) \leq x \right].$$

$\bar{U}_n(x)$ has the same expectation as $U_n(x)$, but its summands are independent. Also each summand lies between 0 and 1, and hence has a variance bounded above by 1/4. Let $X_{(i)}$s denote the ordered $X_i$s and $Y_{[i](j)}$s be their ordered concomitants, i.e.,$Y_{[i](j)}$s are the replications at $X_{(i)}$s and $Y_{[i](j)} \leq Y_{[i](j+1)}, j = 1, \ldots (m-1)$. It can be seen that

$$U_n(x) = E\left[\bar{U}_n(x) | X_{(i)}, Y_{[i](j)}, 1 \leq i \leq n, 1 \leq j \leq m\right].$$

So, by the Rao-Blackwell theorem, $\mathrm{Var}(U_n(x)) \leq \mathrm{Var}(\bar{U}_n(x)) \leq 1/(4s_n) \to 0$ as $s_n = \lfloor n/q_n \rfloor \to \infty$ and thus $U_n(x, \beta) \xrightarrow{P} G_\beta(x)$ for $x \in \mathbb{R}$. The uniform convergence in probability and $(ii)$ follow from arguments for Theorem 15.7.1 in Lehmann and

Romano (2005), given the pointwise convergence shown above. $\square$

## B.7 Proof of Proposition 3.15

We first justify that the rate of convergence of $\hat{d}_{m,n}$ remains unchanged when we impute a $\sqrt{mn}$-consistent estimator of $\tau_0$. Recall that

$$\mathbb{M}_{m,n}(d,\sigma,\tau) = \mathbb{P}_n\left[\left\{\Phi\left(\frac{\sqrt{m}(\bar{Y}-\tau)}{\sigma}\right) - \gamma\right\} 1(X \le d)\right].$$

As in the proof of Theorem 3.2, we need to bound the expression

$$E^* \sup_{\substack{\rho_n(d,d_{m,n})<\delta \\ (\sigma,\tau)\in V_{n,\epsilon}}} \sqrt{n}|\{\mathbb{M}_{m,n}(d,\sigma,\tau)-M_{m,n}(d,\sigma_0,\tau_0)\}-\{\mathbb{M}_{m,n}(d_{m,n},\sigma,\tau)-M_{m,n}(d_{m,n},\sigma_0,\tau_0)\}|,$$

where $M_{m,n}(d,\sigma,\tau) = E[\mathbb{M}_{m,n}(d,\sigma,\tau)]$, and $V_{n,\epsilon} = [\sigma_0 - L_\epsilon/\sqrt{mn}, \sigma_0 + L_\epsilon/\sqrt{mn}] \times [\tau_0 - L_\epsilon/\sqrt{mn}, \tau_0 + L_\epsilon/\sqrt{mn}]$ is a set with $L_\epsilon$ chosen in such a way that $P[(\hat{\sigma},\tau) \in V_{n,\epsilon}] > 1-\epsilon$, for $\epsilon > 0$. Following the proof of Theorem 3.2, the above display can be bounded by

$$\begin{aligned}
&E^* \sup_{\substack{\rho_n(d,d_{m,n})<\delta \\ (\sigma,\tau)\in V_{n,\epsilon}}} \left|\mathbb{G}_n\left[\left\{\Phi\left(\sqrt{m}(\bar{Y}-\tau)/\sigma\right)-\gamma\right\}\left\{1(X\le d)-1(X\le d_{m,n})\right\}\right]\right| \\
&\quad +\sqrt{n} \sup_{\substack{\rho_n(d,d_{m,n})<\delta \\ \sqrt{mn}|\sigma-\sigma_0|<L_\epsilon}} \left|P_n\left[\left\{\Phi\left(\sqrt{m}(\bar{Y}-\tau_0)/\sigma\right)-\Phi\left(\sqrt{m}(\bar{Y}-\tau_0)/\sigma_0\right)\right\}\times\right.\right. \\
&\quad \left.\left.\{1(X\le d)-1(X\le d_{m,n})\}\right]\right| \\
&\quad +\sqrt{n} \sup_{\substack{\rho_n(d,d_{m,n})<\delta \\ (\sigma,\tau)\in V_{n,\epsilon}}} \left|P_n\left[\left\{\Phi\left(\sqrt{m}(\bar{Y}-\tau)/\sigma\right)-\Phi\left(\sqrt{m}(\bar{Y}-\tau_0)/\sigma\right)\right\}\times\right.\right. \\
&\quad \left.\left.\{1(X\le d)-1(X\le d_{m,n})\}\right]\right|.
\end{aligned}$$

The first term involves empirical process acting on a class of functions with VC-index at most 3 while the second term appears in the proof of Theorem 3.2. These two terms can be dealt in the same manner as in that proof. For the third term, note that $|\Phi\left(\sqrt{m}(\bar{Y}-\tau)/\sigma\right)-\Phi\left(\sqrt{m}(\bar{Y}-\tau_0)/\sigma\right)| \le \sup_u |\Phi\left(u+\sqrt{m}(\tau_0-\tau)/\sigma\right)-\Phi\left(u\right)|$ which equals $|\Phi\left(\sqrt{m}(\tau_0-\tau)/2\sigma\right)-\Phi\left(-\sqrt{m}(\tau_0-\tau)/2\sigma\right)|$. As $\Phi$ is Lipschitz of order 1, this is further bounded above by $\sqrt{m}|\tau_0-\tau|/\sigma$. Hence, for sufficiently large $n$, the third term in the above display is bounded by

$$2(L_\epsilon/\sigma_0) \sup_{\rho_n(d,d_{m,n})<\delta} P_n|1(X \le d) - 1(X \le d_{m,n})|.$$

Hence, this term has the same order as $\phi_n(\cdot)$ appearing in (B.2), in the proof of Theorem 3.2. The rest of the argument is identical to the proof for the known $\tau_0$ case and thus, we end up with the same rate of convergence.

To justify that the limiting distributions also stay the same, note that $n(\hat{d}_{m,n}-d_0)$ is a minimizer of the process $n\{\mathbb{M}_n(d_0+t/n,\hat{\sigma},\hat{\tau})-\mathbb{M}_n(d_0,\hat{\sigma},\tau_0)\}$, $t \in \mathbb{R}$. But by arguments analogous to the proof of Lemma B.3, the difference $\sup_{t\in[-L,L]} n|\mathbb{M}_n(d_0+t/n,\hat{\sigma},\hat{\tau})-\mathbb{M}_n(d_0+t/n,\hat{\sigma},\tau_0)|$ is $\sqrt{m}(\hat{\tau}-\tau_0)/\hat{\sigma}\times O_P(1)$, which goes in probability to zero for any $L>0$. Hence, the limiting distribution is not affected as long as we have a $\sqrt{mn}$-consistent estimate of $\tau_0$. □

## B.8 Proof of Proposition 3.16

For notational convenience, we denote $\mathbb{M}_{m,n}^{FD}(d)$ by $\mathbb{M}_n^{FD}(d)$ (as $m$ is a function of $n$). Let $\Phi_n$ be as defined in Section 3.2.1 and

$$M_n^{FD}(d) = E\left[\mathbb{M}_n^{FD}(d)\right] = \frac{1}{n}\sum_{i=1}^{n}\left\{\Phi_n\left(\frac{\sqrt{m}\;\mu(i/n)}{\sqrt{1+\sigma_0^2}}\right)-\gamma\right\}1\left(i/n \le d\right).$$

The expression on the right side follows from calculations almost identical to (3.5). Let $d_n^{FD} = \operatorname{sargmin}_{d\in[0,1]} M_n^{FD}(d)$. To prove Proposition 3.16, we use Theorem 3.2.5 of van der Vaart and Wellner (1996) (see also Theorem 3.4.1) which requires coming up with a non-negative map $d \mapsto \rho_n(d, d_n^{FD})$ such that

$$M_n^{FD}(d) - M_n^{FD}(d_n^{FD}) \gtrsim \rho_n^2(d, d_n^{FD}).$$

Then a bound on the modulus of continuity with respect to $\rho_n$ is needed, i.e.,

$$E\left[\sqrt{n} \sup_{\rho_n(d,d_n^{FD})<\delta} |(\mathbb{M}_n^{FD} - M_n^{FD})(d) - (\mathbb{M}_n^{FD} - M_n^{FD})(d_n^{FD})|\right] \lesssim \phi_n(\delta),$$

where the map $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$. The rate of convergence is then governed by the behavior of $\phi_n$. We start with the following choice for $\rho_n$.

**Lemma B.6.** *Let $\eta > 0$. Let $d \mapsto \rho_n(d, d_n^{FD})$ be a map from $(0,1)$ to $[0,\infty)$ such that*

$$\begin{aligned}\rho_n^2(d, d_n^{FD}) &= (1/n)\{|\lfloor nd \rfloor - \lfloor nd_0 \rfloor| 1(d \leq d_0) \\ &\quad + |\lfloor nd \rfloor - \lfloor n(d_n^{FD} + \eta m^{-1/(2k)}) \rfloor| 1(d > d_n^{FD} + \eta m^{-1/(2k)})\}.\end{aligned}$$

*Then $\eta$ and $\kappa > 0$ can be chosen such that for sufficiently large $n$ and $\rho_n(d, d_n^{FD}) < \kappa$, we have*

$$M_n(d) - M_n(d_n^{FD}) \gtrsim \rho_n^2(d, d_n^{FD}).$$

*Also, $(d_n^{FD} - d_0) = O(m^{-1/(2k)})$.*

We first provide the proof of Proposition 3.16 using Lemma B.6. Using the above lemma, there exists $A < \infty$ such that for sufficiently large $n$ and any $\delta > 0$, $\{\rho_n(d, d_n^{FD}) < \delta\} \subset \{|d - d_n^{FD}| < A(\delta^2 + n^{-\alpha})\}$. Consider the case $d > d_n^{FD}$ and let

$$U(i,d) = \left\{\Phi\left(\sqrt{m}\bar{Y}_i\right) - \Phi_n\left(\frac{\sqrt{m}\mu(i/n)}{\sqrt{1+\sigma_0^2}}\right)\right\} 1\left(d_n^{FD} < i/n \leq d\right).$$

Note that $E\{U(i,d)\} = 0$ and for $1 \leq i \neq j \leq n$, $U(i,d)$ and $U(j,d)$ are independent. Also, $S(i,d) := (\mathbb{M}_n^{FD} - M_n^{FD})(d) - (\mathbb{M}_n^{FD} - M_n^{FD})(d_n^{FD}) = (1/n)\sum_i U(i,d)$, a normalized sum of $(\lfloor nd \rfloor - \lfloor nd_n^{FD} \rfloor)$ non-zero independent terms, is a martingale in $d$, $d \geq d_n^{FD}$, with right continuous paths. As $|U(\cdot, d)| \leq 1$, $E\{U^2(\cdot, d)\}$ is at most 1. Using Doob's inequality, we get

$$
\begin{aligned}
E & \left[ \sup_{0 \leq d - d_n^{FD} < A(\delta^2 + n^{-\alpha})} \sqrt{n} |S(i, d)| \right] \\
& \leq \sqrt{n} \left\{ ES^2(i, d_n^{FD} + A(\delta^2 + n^{-\alpha})) \right\}^{1/2} \\
& = \frac{1}{\sqrt{n}} \left[ \sum_{i \leq n} E\{U^2(i, d_n^{FD} + A(\delta^2 + n^{-\alpha}))\} \right]^{1/2} \\
& \lesssim (\delta^2 + n^{-\alpha})^{1/2}.
\end{aligned}
$$

A similar bound can be established for the case $d \leq d_n^{FD}$. Hence, we get

$$
E \left[ \sqrt{n} \sup_{\rho_n(d, d_n^{FD}) < \delta} |(\mathbb{M}_n^{FD} - M_n^{FD})(d) - (\mathbb{M}_n^{FD} - M_n^{FD})(d_n^{FD})| \right] \lesssim \phi_n(\delta),
$$

where $\phi_n(\delta) = (\delta^2 + n^{-\alpha})^{1/2}$. The function $\phi_n(\cdot)$ and $\rho_n(\cdot, d_n^{FD})$ satisfy the conditions of Theorem 3.2.5 of van der Vaart and Wellner (1996). Hence, the rate of convergence, say $r_n$, satisfies

$$
r_n^2 \phi_n \left( \frac{1}{r_n} \right) \lesssim \sqrt{n} \ \Rightarrow (r_n^2 + r_n^4 n^{-\alpha}) \lesssim n.
$$

Note that $r_n^2 = n^\alpha$ satisfies the above relation and therefore $n^\alpha \rho_n^2(\hat{d}_n, d_n^{FD})$ is $O_P(1)$. Consequently, we get $n^\alpha(\hat{d}_n - d_0) = O_P(1)$. □

*Proof of Lemma B.6.* Since $\mu(x) = 0$ for $x \leq d_0$, note that $d_n^{FD} > d_0$ for sufficiently large $n$. As $\Phi_n(0)$ converges to 1/2, it can be seen that for large $n$ and $d \leq d_0$,

$$
\begin{aligned}
M_n(d) - M_n(d_n^{FD}) & \geq M_n(d) - M_n(d_0) \\
& = \sum_{i=1}^{n} \{\gamma - \Phi_n(0)\} 1 \left( d < \frac{i}{n} \leq d_0 \right)
\end{aligned}
$$

$$\geq \frac{1}{2}\left(\gamma - \frac{1}{2}\right)\{\lfloor nd \rfloor - \lfloor nd_0 \rfloor\}/n. \tag{B.11}$$

Next, we show that

$$\Phi_n\left(\frac{\sqrt{m}\mu(d_n^{FD} + \eta/n^\alpha)}{\sqrt{1+\sigma_0^2}}\right) - \gamma > K_0, \tag{B.12}$$

for sufficiently large $n$ and some $K_0 > 0$. It can be shown that $d_n^{FD}$ converges to $d_0$. Hence, $d_n^{FD}$ is not a boundary point of the interval $[1/n, 1]$ for large $n$; it corresponds to a local minimum of $M_n$, i.e,

$$\Phi_n\left(\sqrt{m}\mu(d_n^{FD})/\sqrt{1+\sigma_0^2}\right) \leq \gamma < \Phi_n\left(\sqrt{m}\mu(d_n^{FD} + 1/n)/\sqrt{1+\sigma_0^2}\right).$$

Thus, $\Phi_n(\sqrt{m}\mu(d_n^{FD})/\sqrt{1+\sigma_0^2})$ converges to $\gamma$ and consequently, $\sqrt{m}\mu(d_n^{FD})/\sqrt{1+\sigma_0^2}$ and $m^{-1/(2k)}(d_n - d_0)$ are $O(1)$. Thus, it suffices to show that $\sqrt{m}(\mu(d_n^{FD} + \eta/\nu_n) - \mu(d_n^{FD}))$ is bounded away from zero to justify (B.12). This can be shown in an identical manner as in the proof of Lemma B.2.

Choose $\kappa > 0$ such that $\mu$ is non-decreasing in $(d_0, d_0 + \kappa)$. For sufficiently large $n$, $d_n^{FD} + \eta m^{-1/(2k)} + 1/n < d_0 + \kappa$ and hence,

$$\begin{aligned}
M_n(d) - M_n(d_n^{FD}) &\geq M_n(d) - M_n(d_0 + \eta m^{-1/(2k)}) \\
&\geq \sum_{d_0 + \eta m^{-1/(2k)} \leq i/n \leq d} \left\{\Phi_n\left(\sqrt{m}\mu(i/n)/\sqrt{1+\sigma_0^2}\right) - \gamma\right\} \\
&\geq K_0(\lfloor nd \rfloor - \lfloor n(d_n^{FD} + \eta m^{-1/(2k)}) \rfloor)/n. \qquad \text{(B.13)}
\end{aligned}$$

Using (B.11) and (B.13), we get the result. $\square$

# APPENDIX C

# Proofs for Chapter 4

We start with proving a few auxiliary results that are repeatedly used in the proofs. Recall that $K$ and $\mu$ are Lipschitz continuous of order $\alpha \in (1/2, 1]$. Let $u_n(x,v) = (1/h_n)\mu(v) K((x-v)/h_n)$ for $x \in [0,1]$ and $v \in \mathbb{R}$.

**Lemma C.1.** *For $\bar{\mu}(\cdot)$ as in (4.5), we have*

$$\sup_{x\in[0,1]} \left| \bar{\mu}(x) - \int_0^1 u_n(x,v)dv \right| = O\left(\frac{1}{(nh_n)^\alpha}\right).$$

*Proof.* Note that $\bar{\mu}(x) = (1/n)\sum_i u_n(x, i/n)$ and $u_n(x,v) = 0$ whenever $|x - v| \geq L_0 h_n$. Moreover, the difference between

$$\bar{\mu}(x) - \int_0^1 u_n(x,v)dv$$

and

$$\sum_{\substack{1\leq i\leq n \\ |x-i/n|\leq L_0 h_n}} \int_{i/n}^{(i+1)/n} \{u_n(x,i/n) - u_n(x,v)\}dv$$

is at most

$$\int_{0}^{1/n} |u_n(x,v)|dv + \int_{x-L_0h_n}^{x-L_0h_n+1/n} |u_n(x,v)|dv$$

which is bounded by $(1/n)\sup_{x,v} u_n(x,v) \leq \|\mu\|_\infty \|K\|_\infty/(nh_n)$. Hence,

$$\begin{aligned}
&\left|\bar{\mu}\left(x\right) - \int_0^1 u_n(x,v)dv\right| \\
&\leq\ O\left(\frac{1}{nh_n}\right) + \sum_{\substack{1\leq i\leq n \\ |x-i/n|\leq L_0h_n}} \int_{i/n}^{(i+1)/n} |u_n(x,i/n) - u_n(x,v)|dv.
\end{aligned}$$

For $v_1, v_2 \in \mathbb{R}$, $h_n|u_n(x,v_1) - u_n(x,v_2)| \leq |\mu(v_1) - \mu(v_2)|K\left((x-v_1)/h_n\right) + |\mu\left(v_2\right)|\left|K\left((x-v_1)/h_n\right) - K\left((x-v_2)/h_n\right)\right|$. As $K$ and $\mu$ are Lipschitz continuous of order $\alpha$, $|u_n(x,v_1) - u_n(x,v_2)| \lesssim 1/h_n^{1+\alpha}|v_1 - v_2|^\alpha$. Also, the cardinality of the set $\{i : 1 \leq i \leq n,\ |x - i/n| \leq L_0h_n\}$ is at most $2L_0nh_n + 2$ and therefore, the above display is further bounded (up to a positive constant multiple) by

$$O\left(\frac{1}{nh_n}\right) + \sum_{\substack{1\leq i\leq n \\ |x-i/n|\leq L_0h_n}} \int_{i/n}^{(i+1)/n} \frac{|i/n - v|^\alpha}{h_n^{1+\alpha}}dv \ \leq\ O\left(\frac{1}{nh_n}\right) + \frac{2L_0nh_n + 2}{(\alpha+1)(nh_n)^{1+\alpha}},$$

which is $O\left(1/(nh_n)^\alpha\right)$. Here, the final bound does not depend on $x$ and thus, we get the desired result. □

Note that the above result holds for generic functions $\mu$ and $K$, satisfying assumptions 1(a), 4(c) and 4(d). Letting $\mu(x) \equiv \sigma_0^2$ and substituting $K^2$ for $K$, we get:

**Corollary C.2.** *Let* $z_n(x,v) = (\sigma_0^2/h_n)K^2\left((x-v)/h_n\right)$. *Then,*

$$\sup_{x\in[0,1]}\left|\Sigma_n^2\left(x\right) - \int_0^1 z_n(x,v)dv\right| = O\left(\frac{1}{(nh_n)^\alpha}\right).$$

As a consequence, when $i/n \in [L_0 h_n, 1 - L_0 h_n]$,

$$\begin{aligned}\Sigma_n^2(i/n) &= \int_0^1 z_n(i/n, v)dv + o(1) \\ &= \sigma_0^2 \min_j \int_{(i-n)/(nh_n)}^{i/(nh_n)} K^2(u)du + o(1) \\ &= \sigma_0^2 \bar{K^2} + o(1). \qquad \text{(C.1)}\end{aligned}$$

## C.1 Proof of Theorem 4.3

To prove Theorem 4.3, we use Theorem 3.2.5 of van der Vaart and Wellner (1996) (see also Theorem 3.4.1) which requires coming up with a non-negative map $d \mapsto \rho_n(d, d_n)$ such that

$$M_n(d) - M_n(d_n) \geq \rho_n^2(d, d_n).$$

Then a bound on the modulus of continuity with respect to $\rho_n$ is needed, i.e.,

$$\begin{aligned}&E\left[\sqrt{n} \sup_{\rho_n(d,d_n)<\delta} |(\mathbb{M}_n(d) - \mathbb{M}_n(d_n)) - (M_n(d) - M_n(d_n))|\right] \qquad \text{(C.2)} \\ &= E\left[\sqrt{n} \sup_{\rho_n(d,d_n)<\delta} |(\mathbb{M}_n - M_n)(d) - (\mathbb{M}_n - M_n)(d_n)|\right] \lesssim \phi_n(\delta),\end{aligned}$$

where the map $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$. The rate of convergence is then governed by the behavior of $\phi_n$. We start with the following choice for $\rho_n$.

**Lemma C.3.** *Fix $\eta > 2L_0 > 0$. Let $d \mapsto \rho_n(d, d_n)$ be a map from $(0, 1)$ to $[0, \infty)$ such that*

$$\begin{aligned}\rho_n^2(d, d_n) &= (K_1/n)\left\{|\lfloor nd \rfloor - \lfloor n(d_0 - L_0 h_n) \rfloor| 1(d \leq d_0 - L_0 h_n)\right. \\ &\left.+ |\lfloor nd \rfloor - \lfloor n(d_n + \eta/\nu_n) \rfloor| 1(d > d_n + \eta/\nu_n)\right\},\end{aligned}$$

*for some $K_1 > 0$. Then $K_1$ and $\kappa > 0$ can be chosen such that for sufficiently large $n$ and $\rho_n(d, d_n) < \kappa$, we have*

$$M_n(d) - M_n(d_n) \geq \rho_n^2(d, d_n).$$

We first provide the proof of Theorem 4.3 using Lemma 4.2. By the above Lemma, there exists $A < \infty$ such that for sufficiently large $n$ and any $\delta > 0$, $\{\rho_n(d, d_n) < \delta\} \subset \{|d - d_n| < \delta^2/K_1 + A/\nu_n + 2/n\}$. Let $d > d_n$ and

$$\begin{aligned} U(i,d) &= \left\{\Phi\left(\sqrt{nh_n}(\hat{\mu}(i/n))\right) - \Phi_{i,n}\left(\frac{\sqrt{nh_n}(\bar{\mu}(i/n))}{\sqrt{1+\Sigma_n^2(i/n)}}\right)\right\} \\ &\quad \times 1\left(d_n < \frac{i}{n} \leq d\right) \end{aligned}$$

where $\hat{\mu}$ is defined in (4.5). By (4.6), $E\{U(i,d)\} = 0$. Also, for $1 \leq i, j \leq n$, $U(i,d)$ and $U(j,d)$ are independent whenever $|i - j| \geq 2L_0 n h_n$. Let $j_1^i = i$ and $j_l^i = j_{l-1}^i + \lceil 2L_0 n h_n \rceil$. Then,

$$S(i,d) := (1/n) \sum_{l: j_l^i \leq n} U(j_l^i, d),$$

a sum of at most $\lceil (d - d_n)/(2L_0 h_n) \rceil$ non-zero independent terms, is a martingale in $d$, $d \geq d_n$, with right continuous paths. As $|U(\cdot, d)| \leq 1$, $E\{U^2(\cdot, d)\}$ is at most 1. Using Doob's inequality, we get

$$\begin{aligned} & E\left[\sup_{\substack{|d-d_n|<\delta^2/K_1+A/\nu_n+2/n \\ d \geq d_n}} |S(i,d)|\right] \\ &\quad \leq \left\{ES^2(i, d_n + \delta^2/K_1 + A/\nu_n + 2/n)\right\}^{1/2} \\ &\quad = \frac{1}{n}\left[\sum_{l: j_l^i \leq n} E\{U^2(j_l^i, d_n + \delta^2/K_1 + A/\nu_n + 2/n)\}\right]^{1/2} \end{aligned}$$

$$\leq \frac{1}{L_0 n h_n^{1/2}} (\delta^2/K_1 + A/\nu_n + 2/n)^{1/2}.$$

As $(\mathbb{M}_n - M_n)(d) - (\mathbb{M}_n - M_n)(d_n) = \sum_{i=1}^{\lceil 2L_0 n h_n \rceil - 1} S(i, d)$, for sufficiently large $n$,

$$\begin{aligned} & E\left[\sqrt{n} \sup_{\rho_n(d,d_n)<\delta, d>d_n} |(\mathbb{M}_n - M_n)(d) - (\mathbb{M}_n - M_n)(d_n)|\right] \\ \leq \;& E\left[\sqrt{n} \sup_{\substack{|d-d_n|<\delta^2/K_1+A/\nu_n+2/n \\ d\geq d_n}} |(\mathbb{M}_n - M_n)(d) - (\mathbb{M}_n - M_n)(d_n)|\right] \\ \leq \;& \sqrt{n}(2L_0 n h_n) \frac{1}{L_0 n h_n^{1/2}} (\delta^2/K_1 + A/\nu_n + 2/n)^{1/2} \lesssim \phi_n(\delta), \end{aligned} \tag{C.3}$$

where $\phi_n(\delta) = \sqrt{n h_n}(\delta^2 + \nu_n^{-1} + n^{-1})^{1/2}$. This bound can also be shown to hold when $d \leq d_n$. Also, $\phi_n(\cdot)$ and $\rho_n(\cdot, d_n)$ satisfy the conditions of Theorem 3.2.5 of van der Vaart and Wellner (1996). Hence, the rate of convergence, say $r_n$, satisfies

$$r_n^2 \phi_n\left(\frac{1}{r_n}\right) \lesssim \sqrt{n} \;\Rightarrow n h_n (r_n^2 + r_n^4/\nu_n + r_n^4/n) \lesssim n.$$

Note that $r_n^2 = \nu_n$ satisfies the above relation and therefore $\nu_n \rho_n^2(\hat{d}_n, d_n)$ is $O_p(1)$. Consequently, we get $\nu_n(\hat{d}_n - d_0) = O_p(1)$. □

*Proof of Lemma C.3.* Since $\bar{\mu}(x) = 0$ for $x < d_0 - L_0 h_n$, note that $d_n > d_0 - L_0 h_n$ for sufficiently large $n$. As $\Phi_{i,n}(0)$ converges to $1/2$ uniformly in $i$, it can be seen that for large $n$ and $d \leq d_0 - L_0 h_n$,

$$\begin{aligned} M_n(d) - M_n(d_n) \;&\geq\; M_n(d) - M_n(d_0 - L_0 h_n) \\ &=\; \sum_{i=1}^{n} \{\gamma - \Phi_{i,n}(0)\} 1\left(d < \frac{i}{n} \leq d_0 - L_0 h_n\right) \\ &\geq\; \frac{1}{2}\left(\gamma - \frac{1}{2}\right) \{\lfloor nd \rfloor - \lfloor n(d_0 - L_0 h_n) \rfloor\}/n. \end{aligned} \tag{C.4}$$

Next, we show that

$$\tilde{\Phi}_n\left(\frac{\sqrt{nh_n}\bar{\mu}(d_n+\eta/\nu_n)}{\sqrt{1+\tilde{\Sigma}_n^2}}\right)-\gamma>K_0, \tag{C.5}$$

for sufficiently large $n$ and some $K_0 > 0$. Using (4.7), note that $\tilde{\Phi}_n\left(\sqrt{nh_n}\bar{\mu}(d_n)/\sqrt{1+\tilde{\Sigma}_n^2(d_n)}\right)$ converges to $\gamma$ and consequently, $\sqrt{nh_n}\bar{\mu}(d_n)/\sqrt{1+\tilde{\Sigma}_n^2(d_n)}$ is $O(1)$. As $\Sigma_n^2(d_n)$ is also $O(1)$, it suffices to show that $\sqrt{nh_n}(\bar{\mu}(d_n+\eta/\nu_n)-\bar{\mu}(d_n))$ is bounded away from zero. To show this, note that by Lemma C.1,

$$\begin{aligned}&\sqrt{nh_n}(\bar{\mu}(d_n+\eta/\nu_n)-\bar{\mu}(d_n))\\ &=\int_{-L_0}^{L_0}\sqrt{nh_n}\left\{\mu(d_n+\eta/\nu_n+uh_n)-\mu(d_n+uh_n)\right\}K(u)du+o(1).\end{aligned}$$

Choose $\kappa>0$ such that $\mu$ is non-decreasing in $(d_0,d_0+3\kappa)$. For sufficiently large $n$, $d_n+\eta/\nu_n+L_0h_n<d_0+3\kappa$, and hence, the integrand in the above display is non-negative. With $L_1$ such that $K_{min}=\inf\{K(x):x\in[-L_1,L_1]\}>0$, the above display is bounded from below by

$$2L_1K_{min}\sqrt{nh_n}(\mu(d_n+\eta/\nu_n-L_1h_n)-\mu(d_n+L_0h_n)).$$

As $\eta>2L_0$, note that $d_n+\eta/\nu_n-L_1h_n>d_n+L_0h_n>d_0$. With $\zeta_n^{(1)}$ and $\zeta_n^{(2)}$ being some points in $(d_0,d_n+\eta/\nu_n-L_1h_n)$ and $(d_0,d_n+L_0h_n)$ respectively, we have

$$\begin{aligned}&\sqrt{nh_n}\{\mu(d_n+\eta/\nu_n-L_0h_n)-\mu(d_n+L_0h_n)\}\\ &=\frac{\sqrt{nh_n}}{k!}\{\mu^{(k)}(\zeta_n^{(1)})(d_n+\eta/\nu_n-L_1h_n-d_0)^k-\mu^{(k)}(\zeta_n^{(2)})(d_n+L_0h_n-d_0)^k\}\\ &>\frac{\sqrt{nh_n}\mu^{(k)}(\zeta_n^{(1)})}{k!}[(d_n+\eta/\nu_n-L_1h_n-d_0)^k-(d_n+L_0h_n-d_0)^k]\end{aligned}$$

$$
\begin{aligned}
& + \frac{\sqrt{nh_n}}{k!}[\mu^{(k)}(\zeta_n^{(1)}) - \mu^{(k)}(\zeta_n^{(2)})](d_n + L_0 h_n - d_0)^k \\
> & \sqrt{nh_n}\left[\frac{\left\{\mu^{(k)}(d_0+) + o(1)\right\}(\eta/\nu_n - 2L_0 h_n)^k}{k!} + o(1)(d_n - d_0 + L_0 h_n)^k\right].
\end{aligned}
$$

Using Lemma 4.2, $(d_n - d_0)$ is $O(1/\nu_n)$ and hence, the above display is further bounded from below by

$$
\frac{\sqrt{nh_n}}{\nu_n^k}\left[\frac{\mu^{(k)}(d_0+)}{k!}(\eta - 2L_0)^k + o(1)\right].
$$

As $\sqrt{nh_n}/\nu_n^k \geq 1$, (C.5) holds.

Further, as the kernel $K(u)$ is non-increasing in $|u|$, $\bar{\mu}$ is non-decreasing in $(d_0, d_0 + 2\kappa)$. For $d \in (d_n + \eta/\nu_n, d_0 + 2\kappa)$,

$$
\begin{aligned}
M_n(d) - M_n(d_n) & \geq M_n(d) - M_n(d_0 + \eta/\nu_n) \\
& \geq \sum_{d_0 + \eta/\nu_n \leq i/n \leq d} \left\{\Phi_{i,n}\left(\sqrt{nh_n}\bar{\mu}(i/n)/\sqrt{1 + \Sigma_n^2(i/n)}\right) - \gamma\right\} \\
& \geq K_0(\lfloor nd \rfloor - \lfloor n(d_n + \eta/\nu_n)\rfloor)/n. \qquad \text{(C.6)}
\end{aligned}
$$

Using Lemma 4.2, there exists $A_0 < \infty$ such that for sufficiently large $n$, $\nu_n|d_0 - d_n| \leq A_0$, and hence $\{\rho_n(d, d_n) < \kappa\} \subset \{|d - d_0| < \kappa^2/K_1 + A/\nu_n + 2/n\} \subset \{|d - d_0| < 2\kappa\}$, where $A = 2\max(\eta, L_0, A_0)$. Letting $K_1 = (1/2)\min(\gamma - 1/2, K_0)$ and using (C.4) and (C.6), we get the desired result. □

## C.2 Proof of Lemma 4.4

In order to prove *Lemma 4.4*, we first justify a few auxiliary results required to prove the tightness of $W_n$. Recall that

$$
W_n(t) = \sqrt{nh_n}\hat{\mu}(d_0 + th_n).
$$

Let $\bar{\epsilon}_n(\cdot)$ be such that $\bar{\epsilon}_n(t) = W_n(t) - \sqrt{nh_n}\bar{\mu}(d_0 + th_n)$, i.e.,

$$\bar{\epsilon}_n(t) = \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} \epsilon_i K\left(\frac{d_0 - i/n}{h_n} + t\right). \tag{C.7}$$

**Lemma C.4.** *The processes $\sqrt{nh_n}\bar{\epsilon}_n(d_0 + th_n)$, $t \in \mathbb{R}$, are asymptotically tight in $C(\mathbb{R})$.*

*Proof.* As the kernel $K$ is Lipschitz of order $\alpha > 1/2$, there exists a constant $C_0 > 0$, such that $|K(t) - K(s)| \leq C_0|t-s|^\alpha$. Fix $T > 0$. For $s, t \in [-T, T]$, we have

$$\begin{aligned}
&E\left[\bar{\epsilon}_n(t) - \bar{\epsilon}_n(s)\right]^2 \\
&= \frac{1}{nh_n}\sum_{i=1}^{n} \sigma_0^2 \left|K\left(\frac{d_0 - i/n}{h_n} + t\right) - K\left(\frac{d_0 - i/n}{h_n} + s\right)\right|^2 \\
&= \frac{1}{nh_n}\sum_{i=1}^{n} \sigma_0^2 \left|K\left(\frac{d_0 - i/n}{h_n} + t\right) - K\left(\frac{d_0 - i/n}{h_n} + s\right)\right|^2 \\
&= \frac{1}{nh_n}\sum_{|d_0 - i/n| < (L_0 + T)h_n} \sigma_0^2 \left|K\left(\frac{d_0 - i/n}{h_n} + t\right) - K\left(\frac{d_0 - i/n}{h_n} + s\right)\right|^2 \\
&\leq 4(L_0 + T + 1)\sigma_0^2 C_0^2 |t - s|^{2\alpha}.
\end{aligned}$$

Since $\alpha > 1/2$, the result is a consequence of Theorem 12.3 of Billingsley (1968, pp. 95). $\square$

We use a version of the Arzela-Ascoli theorem to prove the next result and thus we state it below for convenience.

**Theorem C.5** (Arzela-Ascoli)**.** *Let $f_n$ be a sequence of continuous functions defined on a compact set $[a, b]$ such that $f_n$ converge pointwise to $f$ and for any $\delta_n \downarrow 0$ $\sup_{|x-y|<\delta_n} |f_n(x) - f_n(y)|$ converges to 0. Then $\sup_{x\in[a,b]} |f_n(x) - f(x)|$ converges to zero.*

**Lemma C.6.** *The sequence of functions $\sqrt{nh_n}\bar{\mu}(d_0 + th_n)$ converges to $m(t)$, uniformly over compact sets in $\mathbb{R}$.*

*Proof.* The pointwise convergence is evident from Lemma 4.4. To justify the uniform convergence, let $\bar{z}_n(x,t) = (1/h_n)\mu(x)K((d_0 - x)/h_n + t)$. By arguments similar to those for Lemma C.1, $|\bar{z}_n(x,t) - \bar{z}_n(y,t)| \lesssim 1/h_n^{1+\alpha}|x-y|^\alpha$ and consequently, for $t \in [-T,T]$,

$$\begin{aligned}
&\left|\bar{\mu}(d_0 + th_n) - \int \bar{z}_n(x,t)dx\right| \\
&\quad\leq\; O\left(\frac{1}{nh_n}\right) + \sum_{\substack{1\leq i\leq n \\ |d_n - i/n| \leq (L_0+T)h_n}} \frac{|i/n - x|^\alpha}{h_n^{1+\alpha}}dx = O\left(\frac{1}{(nh_n)^\alpha}\right).
\end{aligned}$$

As the above bound does not depend on $t$ and $\alpha > 1/2$, for $s,t \in [-T,T]$, and $\delta > 0$,

$$\begin{aligned}
&\sup_{|t-s|<\delta}\left|\sqrt{nh_n}\bar{\mu}(d_0+th_n) - \sqrt{nh_n}\bar{\mu}(d_0+sh_n)\right| \\
&\quad=\; \sup_{|t-s|<\delta}\left|\sqrt{nh_n}\int_{-\infty}^{\infty}\{\bar{z}_n(x,t) - \bar{z}_n(x,s)\}dx\right| + o(1) \\
&\quad\leq\; \sqrt{nh_n}\int_{-\infty}^{\infty}\mu(d_0+uh_n)\left|K(t-u) - K(s-u)\right|du + o(1) \\
&\quad\leq\; \sqrt{nh_n}\int_{0}^{L_0+T}\frac{\mu^{(k)}(\zeta_u)}{k!}(uh_n)^k\left|K(t-u) - K(s-u)\right|du + o(1),
\end{aligned}$$

where $\zeta_u$ is some intermediate point between $d_0$ and $d_0 + uh_n$. The $k$-th derivative of $\mu$ is bounded on $(d_0, d_0 + (L_0+T)h_n)$ for sufficiently large $n$ and $h_n^k\sqrt{nh_n}$ equals $h_0^{k+1/2}$. As $K$ is uniformly continuous, the above display goes to zero as $\delta \to 0$ by DCT. Hence, by the Arzela–Ascoli theorem we get the desired result. □

We now continue with the proof of Lemma 4.4. For $(a_i, t_i) \in \mathbb{R}^2$, $i = 1, \ldots l$, we have

$$\sum_{i,j} a_i a_j \mathrm{Cov}(W(t_i), W(t_j)) = \int \left\{\sum_i a_i K(t_i + u)\right\}^2 du \geq 0.$$

Hence, the defined covariance function is non-negative definite and by Kolmogorov consistency, the Gaussian process $W$ exists.

Let $r(h) = \left\{\int K(h+u)K(u)du\right\} / \bar{K}^2$ denote the correlation function of $W$. For $W$ to have a continuous modification, by Hunt's theorem (e.g., see Cramér and Leadbetter (1967, pp. 169–171)), it suffices to show that $r(h)$ is $1-O((log(h))^{-\delta})$ for some $\delta > 3$ as $h \to 0$. Note that the kernel $K$ is Lipschitz continuous of order $\alpha$ and hence, we have

$$
\begin{aligned}
|(1-r(h))(log(h))^{\delta}| &= \left| \frac{h^{\alpha}(log(h))^{\delta}}{\int K^2(u)du} \left( \int_{-L_0}^{L_0} \frac{(K(u)-K(h+u))}{h^{\alpha}} K(u)du \right) \right| \\
&\lesssim \frac{1}{\int K^2(u)du} \left| h^{\alpha}(log(h))^{\delta} \right| \to 0.
\end{aligned}
$$

Thus, $W$ has a continuous modification. Next, we justify weak convergence of the process $W_n$ to $W$.

As a consequence of Lemma C.4 and C.6, the process $W_n$ is asymptotically tight. To justify finite dimensional convergence, it suffices to show that:

$$
\begin{pmatrix} W_n(t_1) \\ W_n(t_2) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} W(t_1) \\ W(t_2) \end{pmatrix}, \tag{C.8}
$$

where $t_1, t_2 \in \mathbb{R}$. Let $x_j = d_0 + t_j h_n$, $j = 1$ and 2. Then,

$$
\begin{aligned}
\bar{\mu}(x_j) &= \sqrt{nh_n} \left\{ \int_0^1 \frac{1}{h_n} \mu(x) K\left(\frac{x_j - x}{h_n}\right) dx + O\left(\frac{1}{(nh_n)^{\alpha}}\right) \right\} \\
&= \sqrt{nh_n} \int_{(d_0-1)/h_n+t_j}^{d_0/h_n+t_j} \mu(d_0 + (t_j - v)h_n) K(v)\, dv + o(1) \\
&= \sqrt{nh_n} \int_{(d_0-1)/h_n+t_j}^{t_j} \frac{\mu^{(k)}(d_0^+)}{k!} \left((t_j - v)^k h_n^k + o(h_n^k)\right) K(v)\, dv + o(1)
\end{aligned}
$$

$$= h_0^{k+1/2} \frac{\mu^{(k)}(d_0+)}{k!} \left( \int_{-\infty}^{t_j} (t_j - v)^k K(v)\, dv + o(1) \right) + o(1) = m(t_j) + o(1).$$

The last step follows from DCT as the $k$-th derivative of $\mu$ is bounded in a right neighborhood of $d_0$ and $\int |v|^k K(v) dv < \infty$. Moreover,

$$\begin{aligned} E[\bar{\epsilon}_n(x_j)] &= 0, \\ Var[\bar{\epsilon}_n(x_j)] &= \Sigma_n^2(x_j) \to \sigma_0^2 \bar{K}^2, \end{aligned}$$

and, by a change of variable,

$$\begin{aligned} &\mathrm{Cov}[\bar{\epsilon}_n(x_1), \bar{\epsilon}_n(x_2)] \\ &= \mathrm{Cov}\left[ \frac{1}{\sqrt{nh_n}} \frac{\sum \epsilon_i K((x_1 - i/n)/h_n)}{\Sigma_n(x_1)}, \frac{1}{\sqrt{nh_n}} \frac{\sum \epsilon_i K((x_2 - i/n)/h_n)}{\Sigma_n(x_2)} \right] \\ &= \sigma_0^2 \int K(t_1 + u) K(t_2 + u) du + o(1). \end{aligned}$$

Also,

$$\frac{\max_i K^2((x_j - i/n)/h_n)}{\sum K^2((x_j - i/n)/h_n)} \leq \frac{\|K\|_\infty^2}{nh_n(\bar{K}^2 + o(1))} \to 0.$$

Hence, the Lindeberg–Feller condition is satisfied for $\bar{\epsilon}_n(x_j)$s and by the Cramér-Wold device, (C.8) holds. This justifies the finite dimensional convergence and hence, we have the result. □

## C.3 Proof of Theorem 4.6

In order to prove Theorem 4.6, an ergodic theorem and Borell's inequality are found useful, which are stated below for convenience. For the proofs of the two results, see, for example, Cramér and Leadbetter (1967, pp. 147), and (Adler and Taylor, 2007, pp. 49–53), respectively. Also, we use Theorem 3 from Ferger (2004) which we state below as well.

**Theorem C.7.** *Consider a real continuous second order stationary process $\xi(t)$ with mean 0 and correlation function $R(t)$. If*

$$\frac{1}{T}\int_0^T R(t)dt = O\left(\frac{1}{T^a}\right)$$

*for any $a > 0$, then $\xi$ satisfies the law of large numbers, i.e., $T^{-1}\int_0^T \xi(t)dt$ converges a.s. to zero as $T \to \infty$.*

**Theorem C.8** (Borell's inequality)**.** *Let $\xi$ be a centered Gaussian process, a.s. bounded on a set $I$. Then $E\left\{\sup_{u\in I}\xi(u)\right\} < \infty$ and for all $x > 0$,*

$$P\left\{\sup_{u\in I}\xi(u) - E\left(\sup_{u\in I}\xi(u)\right) > x\right\} \leq \exp\left(\frac{-x^2}{2\sigma_I^2}\right),$$

*where $\sigma_I^2 = \sup_{u\in I} Var\{\xi(u)\}$.*

**Theorem C.9** (Ferger (2004))**.** *Let $\mathbb{V}_n$, $n \geq 0$, be stochastic processes in $D(\mathbb{R})$, defined on a common probability space $(\Omega, \mathcal{A}, P)$. Let $\xi_n$ be a Borel-measurable minimizer of $\mathbb{V}_n$. Suppose that:*

*(i)* *$\mathbb{V}_n$ converges weakly to $\mathbb{V}_0$ in $D[-C, C]$ for each $C > 0$.*

*(ii)* *The trajectories of $\mathbb{V}_0$ almost surely possess a smallest and a largest minimizer $\xi_0^s$ and $\xi_0^l$ respectively, which are Borel measurable.*

*(iii)* *The sequence $\xi_n$ is uniformly tight.*

*Then for every $x \in \mathbb{X}$,*

$$P[\xi_0^l < x] \leq \liminf_{n\to\infty} P_*[\xi_n < x] \leq \limsup_{n\to\infty} P^*[\xi_n \leq x] \leq P[\xi_0^s \leq x].$$

*Here, $\mathbb{X} = \{x \in \mathbb{R} : P[\mathbb{V}_0$ is continuous at $x] = 1\}$.*

We now continue with the proof of Theorem 4.6. Let

$$W_0(t) = W(t) - m(t).$$

This is a mean zero stationary process and thus, so is the process $D(t) = \Phi(W_0(t)) - 1/2$ with correlation function $R(t)$, say. As, $K$ is supported on $[-L_0, L_0]$, $W(t_1)$ and $W(t_2)$ are independent whenever $|t_1 - t_2| \geq 2L_0$ and hence $R(t) = 0$ for $t > 2L_0$. So, $(1/t)\int_0^t R(y)dy = O(1/t)$ as $|t| \to \infty$ and therefore, by *Theorem C.7*, $Z_1(t) = (1/t)\int_0^t D(y)dy \to 0$ a.s. as $|t| \to \infty$. For $t < 0$, we write $Z(t)$ as

$$Z(t) \;=\; t\left[Z_1(t) + (1/2 - \gamma) + (1/t)\int_0^t \left\{\Phi\left(W(t)\right) - \Phi\left(W_0(t)\right)\right\} dy\right].$$

When $t < -L_0$, $m(t) = 0$, which gives $W(t) = W_0(t)$ and hence the third term in the above display goes to zero and $Z(t) \to \infty$ a.s. as $t \to -\infty$. For $t > 0$, fix $M > 0$ and $j$ be a positive integer. Then

$$\begin{aligned}
P\left[\inf_{t\in[j,j+1]} W(t) < M\right] &\leq P\left[\inf_{t\in[j,j+1]} W_0(t) + \inf_{t\in[j,j+1]} m(t) < M\right] \\
&= P\left[\sup_{t\in[j,j+1]} (-W_0(t)) > m(j) - M\right],
\end{aligned}$$

as $\inf_{t\in[j,j+1]} m(t) = m(j)$. By Borell's inequality, the above probability is bounded by $\exp\left[\left\{-m(n) - L_0 - E\sup_{t\in[j,j+1]}(-W_0(t))\right\}^2\right]$, where by stationarity, $E\sup_{t\in[j,j+1]}(-W_0(t)) = E\sup_{t\in[0,1]}(-W_0(t))$ which is finite, again due to Borell's inequality. Also, it can be seen that $m(j) \gtrsim (j - L_0)^k$ and hence $\sum_{j\geq 1} P\left[\sup_{t\in[j,j+1]}(-W_0(t)) > m(j) - M\right] < \infty$. Using Borel–Cantelli lemma, we get $P[\liminf_{t\to\infty} W(t) > M] = 1$. As $M$ can be made arbitrarily large, we get that $W(t)$ diverges to $\infty$ a.s. as $t \to \infty$ and consequently so does $Z(t)$.

Note that $Z_n$ (defined in (4.9)) converges weakly to $Z$ in $B_{loc}(\mathbb{R})$ and consequently, in $D(\mathbb{R})$ as well. Moreover, $Z$ has continuous sample paths with probability 1. As $Z(t) \to \infty$ when $|t| \to \infty$, $\xi_0^s$ and $\xi_0^l$ are well defined and Borel measurable. Further, recall that $h_n^{-1}(\hat{d}_n - d_0)$, the smallest argmin of the process $Z_n(\cdot)$, is determined by the ordering of finitely many random variables and hence, is measurable. Also, by Theorem 4.3, it is $O_p(1)$. Hence, conditions (i), (ii) and (iii) of Theorem C.9 are satisfied with $\mathbb{V}_n = Z_n$ and $\mathbb{V}_0 = Z$, and thus,

$$\begin{aligned}
\liminf_{n\to\infty} P[c^s_{\alpha/2} < h_n^{-1}(\hat{d}_n - d_0) < c^l_{1-\alpha/2}] &\geq \liminf_{n\to\infty} P[h_n^{-1}(\hat{d}_n - d_0) < c^l_{1-\alpha/2}] \\
&\quad - \limsup_{n\to\infty} P[h_n^{-1}(\hat{d}_n - d_0) \leq c^s_{\alpha/2}] \\
&\geq 1 - \alpha.
\end{aligned}$$

Hence, we get the desired result. □

## C.4 Outline of the proof of Proposition 4.9

We assume the rate of convergence for the proof as it is a consequence of arguments similar to that for the proof of Proposition 4.11 (see Section C.6).

To see that we end up with the given limiting distribution, recall that for $\tau_0 = 0$,

$$\hat{d}_n^1 = \operatorname*{sargmin}_{d\in[0,1]} \frac{1}{n}\sum_{i=1}^{n}\left\{\Phi\left(\frac{\sqrt{nh_n}\hat{\mu}(i/n)}{\Sigma_n(i/n,\hat{\sigma})}\right) - \gamma\right\} 1\left(\frac{i}{n} \leq d\right).$$

Thus, the form of the limit distribution is dictated by the asymptotic behavior of the local process

$$Z_n^1(t) = \frac{1}{nh_n}\sum_{i=1}^{n}\left\{\Phi\left(\frac{\sqrt{nh_n}\hat{\mu}(i/n)}{\Sigma_n(i/n,\hat{\sigma})}\right) - \gamma\right\}\left(1\left(\frac{i}{n} \leq d_0 + th_n\right) - 1\left(\frac{i}{n} \leq d_0\right)\right).$$

Proceeding as we did in the proof of Theorem 4.5, $Z_n^1$ can be split into $I_n^1(t) + II_n^1(t)$, where

$$II_n^1(t) = h_n^{-1} \int_{d_0}^{d_0+th_n} \left( \Phi \left( \frac{\sqrt{nh_n}\hat{\mu}(x)}{\Sigma_n(x,\hat{\sigma})} \right) - \gamma \right) dx \tag{C.9}$$

and the contribution of $I_n^1(t) = Z_n^1(t) - II_n^1(t)$ can be shown to converge to zero. By a change of variable, $II_n^1$ can be written as

$$II_n^1(t) = \int_0^t \left[ \Phi \left( \frac{W_n(y)}{\Sigma_n(d_0 + yh_n, \hat{\sigma})} \right) - \gamma \right] f(d_0 + yh_n) dy,$$

where, $W_n$ is as defined in (4.10). This term differs from its analogue for **Method 2** (see (4.3)) through the normalizing factor $\Sigma_n(d_0 + yh_n, \hat{\sigma})$ which converges in probability to $\sigma_0\sqrt{\bar{K^2}}$. The tightness of the ratio process $W_n(y)/\Sigma_n(d_0 + yh_n, \hat{\sigma})$ can be established through calculations similar to those in the proof of Lemma 4.4. Hence, by a Slutsky-type argument, we get that

$$h_n^{-1}(\hat{d}_n^1 - d_0) \xrightarrow{d} \underset{t\in\mathbb{R}}{\operatorname{argmin}} \int_0^t \left\{ \Phi \left( \frac{W(y)}{\sigma_0\sqrt{\bar{K^2}}} \right) - \gamma \right\} dy,$$

for $h_n = h_0 n^{-1/(2k+1)}$. Note that the process on the right side of the above display is precisely $Z_1$. This completes the proof. $\square$

## C.5 Outline of the proof of Proposition 4.10

Here, we provide a brief outline of the proof to convince the reader about the form of the limiting distribution. Note that this is dictated by the asymptotic behavior of the local process

$$\tilde{Z}_n(t) \quad = \quad \mathbb{P}_n \left[ \left\{ \Phi \left( \sqrt{nh_n}\tilde{\mu}(X) \right) - \gamma \right\} (1(X \le d_0 + th_n) - 1(X \le d_0)) \right]$$

that arises out of the criterion in (4.11) (with $\tau_0 = 0$). As in the proof of Theorem 4.5, $\tilde{Z}_n$ can be split into $\tilde{I}_n(t) + \tilde{II}_n(t)$, where

$$\tilde{II}_n(t) = h_n^{-1} \int_{d_0}^{d_0+th_n} \left(\Phi\left(\sqrt{nh_n}\tilde{\mu}(x)\right) - \gamma\right) f(x)dx \tag{C.10}$$

and the contribution of $\tilde{I}_n(t) = \tilde{Z}_n(t) - \tilde{II}_n(t)$ can be shown to go to zero. By a change of variable, $\tilde{II}_n$ can be written as

$$\tilde{II}_n(t) = \int_0^t \left[\Phi\left(\tilde{W}_n(y)\right) - \gamma\right] f(d_0 + yh_n)dy,$$

where $\tilde{W}_n(y) = \sqrt{nh_n}\tilde{\mu}(d_0 + yh_n)$. The process $\tilde{W}_n$ can be shown to converge weakly to the process $\tilde{W}$ by an imitation of the arguments in the proof of Lemma 4.4. Also, $f(d_0 + yh_n)$ converges to $f(d_0) > 0$. Consequently

$$h_n^{-1}(\tilde{d}_n - d_0) \xrightarrow{d} \operatorname*{argmin}_{t\in\mathbb{R}} \left\{f(d_0)\tilde{Z}(t)\right\} = \operatorname*{argmin}_{t\in\mathbb{R}} \left\{\tilde{Z}(t)\right\}.$$

## C.6 Proof of Proposition 4.11

Recall that

$$\mathbb{M}_n(d, \tilde{\tau}) = \frac{1}{n}\sum_{i=1}^{n} \left[\Phi\left(\sqrt{nh_n}\left(\hat{\mu}\left(\frac{i}{n}\right) - \tilde{\tau}\right)\right) - \gamma\right] 1\left(\frac{i}{n} \le d\right)$$

and $M_n(d, \tilde{\tau}) = E[\mathbb{M}_n(d, \tilde{\tau})]$. We make the dependence on the parameter $\tau$ explicit for the analysis. Here $M_n(d, \hat{\tau})$ is interpreted as $M_n(d, \tilde{\tau})$ computed at $\tilde{\tau} = \hat{\tau}$. Now, we extend the proof of Theorem 4.3 to show that the rate of convergence remains the same.

*Rate of convergence.* As $\sqrt{n}(\hat{\tau} - \tau_0) = O_p(1)$, for any $\epsilon > 0$, there exists $V_{\epsilon/2} > 0$ such that $P\left[\sqrt{n}|\hat{\tau} - \tau_0| < V_{\epsilon/2}\right] > 1 - \epsilon$. To show that the rate of convergence does not change, we need to derive a bound on

$$E\left[\sqrt{n}\sup_{\rho_n(d,d_n)<\delta}|(\mathbb{M}_n(d,\hat{\tau})-\mathbb{M}_n(d,\hat{\tau}))-(M_n(d_n,\tau_0)-M_n(d_n,\tau_0))|\right]$$

having the same order as $\phi_n(\delta)$ (see (C.2) and (C.3)). A relaxation is possible due to Theorem B.1 of Appendix B. For each $\epsilon>0$, it suffices to find a bound of the form

$$E\left[\sqrt{n}\sup_{\rho_n(d,d_n)<\delta}|(\mathbb{M}_n-M_n)(d,\hat{\tau})-(\mathbb{M}_n-M_n)(d_n,\tau_0)|1(U_{n,\epsilon})\right]\leq C_\epsilon\phi_n(\delta), \quad \text{(C.11)}$$

where $P[U_{n,\epsilon}]>1-\epsilon$ and $C_\epsilon>0$; see Banerjee and McKeague (2007, Theorem 5.2). For $U_{n,\epsilon}=[\hat{\tau}\in[\tau_0-V_{\epsilon/2}/\sqrt{n},\tau_0+V_{\epsilon/2}/\sqrt{n}]]$, the left side of the above display can be bounded by

$$\begin{aligned}
&E\left[\sqrt{n}\sup_{\rho_n(d,d_n)<\delta,|\tilde{\tau}-\tau_0|<V_{\epsilon/2}/\sqrt{n}}|(\mathbb{M}_n-M_n)(d,\tilde{\tau})-(\mathbb{M}_n-M_n)(d_n,\tau_0)|\right]\\
\leq\;& E\left[\sqrt{n}\sup_{\rho_n(d,d_n)<\delta}|(\mathbb{M}_n-M_n)(d,\tau_0)-(\mathbb{M}_n-M_n)(d_n,\tau_0)|\right]\\
&+E\left[\sqrt{n}\sup_{\substack{\rho_n(d,d_n)<\delta,\\ |\tilde{\tau}-\tau_0|<V_{\epsilon/2}/\sqrt{n}}}|(\mathbb{M}_n(d,\tilde{\tau})-\mathbb{M}_n(d,\tau_0))-(\mathbb{M}_n(d_n,\tilde{\tau})-\mathbb{M}_n(d_n,\tau_0))|\right].
\end{aligned}$$

The first term on the right side is precisely the term dealt in the case of a known $\tau_0$ (see (C.2)). As for the second term, note that by the Lipschitz continuity of $\Phi$,

$$\begin{aligned}
&|(\mathbb{M}_n(d,\tilde{\tau})-\mathbb{M}_n(d,\tau_0))-(\mathbb{M}_n(d_n,\tilde{\tau})-\mathbb{M}_n(d_n,\tau_0))|\\
\leq\;& \frac{1}{n}\sum_{i=1}^{n}\left\{\left|\Phi\left(\sqrt{nh_n}\left(\hat{\mu}\left(\frac{i}{n}\right)-\tilde{\tau}\right)\right)-\Phi\left(\sqrt{nh_n}\left(\hat{\mu}\left(\frac{i}{n}\right)-\tau_0\right)\right)\right|\right.\\
&\left.\times\left|1\left(\frac{i}{n}\leq d\right)-1\left(\frac{i}{n}\leq d_n\right)\right|\right\}\\
\lesssim\;& \sqrt{nh_n}|\tilde{\tau}-\tau_0|\frac{|\lceil nd\rceil-\lceil nd_n\rceil|}{n}.
\end{aligned}$$

Thus, we have

$$E\left[\sqrt{n}\sup_{\substack{\rho_n(d,d_n)<\delta,\\ |\tilde{\tau}-\tau_0|<V_{\epsilon/2}/\sqrt{n}}}|(\mathbb{M}_n(d,\tilde{\tau})-\mathbb{M}_n(d,\tau_0))-(\mathbb{M}_n(d_n,\tilde{\tau})-\mathbb{M}_n(d_n,\tau_0))|\right]$$
$$\lesssim \sqrt{n}\frac{\sqrt{nh_n}V_{\epsilon/2}}{\sqrt{n}}(\delta^2+2/n)\lesssim V_{\epsilon/2}\phi_n(\delta),$$

for $\delta<1$ and large $n$. Hence, the expression in (C.11) has the same bound $\phi_n(\cdot)$ (up to a different constant) and thus, we get the same rate of convergence.

*Limit distribution.* Recall from (4.9) that

$$Z_n(t)=Z_n(t,\tau_0)=h_n^{-1}\left[\mathbb{M}_n(d_0+th_n,\tau_0)-\mathbb{M}_n(d_0,\tau_0)\right].$$

To show that the limiting distribution of $\hat{d}_n$ remains the same, it suffices to show that

$$\sup_{t\in[-T,T]}|Z_n(t,\hat{\tau})-Z_n(t,\tau_0)| \tag{C.12}$$

converges in probability to zero, for any $T>0$. Again by the Lipschitz continuity of $\Phi$,

$$\begin{aligned}
&|Z_n(t,\hat{\tau})-Z_n(t,\tau_0)|\\
&= \frac{1}{nh_n}\left|\sum_{i=1}^{n}\left\{\Phi\left(\sqrt{nh_n}(\hat{\mu}\left(\frac{i}{n}\right)-\hat{\tau})\right)-\Phi\left(\sqrt{nh_n}(\hat{\mu}\left(\frac{i}{n}\right)-\tau_0)\right)\right.\right.\\
&\quad\left.\left.\times\left(1\left(\frac{i}{n}\le d_0+th_n\right)-1\left(\frac{i}{n}\le d_0\right)\right)\right\}\right|\\
&\lesssim \frac{1}{nh_n}\left|\sum_{i=1}^{n}\sqrt{nh_n}|\hat{\tau}-\tau_0|\left(1\left(\frac{i}{n}\le d_0+Th_n\right)-1\left(\frac{i}{n}\le d_0\right)\right)\right|\\
&\le h_n\frac{(Tnh_n+2)}{nh_n}\sqrt{n}|\hat{\tau}-\tau_0|.
\end{aligned}$$

As the above bound is uniform in $t \in [-T, T]$ and $\sqrt{n}(\hat{\tau} - \tau_0)$ is $O_p(1)$, the expression in (C.12) converges in probability to zero and hence, we get the desired result. $\square$

## C.7 Proof of Proposition 4.13

Given what has been done earlier for proving results from Section 4.2.2, it suffices to show that the process $\bar{\epsilon}_n(t)$, defined in (C.7), converges weakly to a mean zero Gaussian process having the covariance function of $\bar{W}$ in the setup of Section 4.5. As $W_n(t) = \sqrt{nh_n}\bar{\mu}(d_0 + th_n) + \bar{\epsilon}_n(t)$, Lemma C.6 then justifies the weak convergence of $W_n$ to $\bar{W}$. The statement and the proof of Lemma 4.4 relies on the i.i.d. assumption only through the convergence of $W_n$'s and the form of their limit. Hence, it would follow that the process $Z_n$ (defined in (4.9)) converges to $\bar{Z}$. The result then follows from applying the argmin continuous mapping theorem as in proving Theorem 4.7.

We start by showing the covariance function of the process $\bar{\epsilon}_n$ converges to that of $\bar{W}$. For $t_1, t_2 \in \mathbb{R}$, let $x_j = d_0 + t_j h_n$, $j = 1, 2$. We have

$$\mathrm{Cov}(\bar{\epsilon}_n(t_1), \bar{\epsilon}_n(t_2)) \quad = \quad \frac{\sigma_0^2}{nh_n} \sum_{l,j} \rho(l-j) K\left(\frac{x_1 - l/n}{h_n}\right) K\left(\frac{x_2 - j/n}{h_n}\right).$$

As $\sigma_0^2 \rho(l-j) = \int_{-\pi}^{\pi} \psi(u) \exp(\imath(l-j)u) du$, the above expression reduces to

$$\frac{1}{nh_n} \int_{-\pi}^{\pi} \psi(u) \hat{K}_{x_1}(u) \hat{K}_{x_2}(-u) du,$$

where for $x, u \in \mathbb{R}$, $\hat{K}_x(u) = \sum_j K(h^{-1}\{x - j/n\}) e^{\imath j u}$. Under short range dependence, Assumption 1 of Robinson (1997) requires $\psi$ to be an even non-negative function which is continuous and positive at 0. Using this assumption, it can be shown that the difference between the above display and

$$\frac{\psi(0)}{nh_n}\int_{-\pi}^{\pi}\hat{K}_{x_1}(u)\hat{K}_{x_2}(-u)du$$

goes to zero by calculations almost identical to those in Robinson (1997, pp. 2061–2062). As $\int_{-\pi}^{\pi}\exp(\imath(l-j)u)du = 2\pi\delta_{lj}$, with $\delta_{lj}$ being the Kronecker delta, the above expression equals

$$\frac{2\pi\psi(0)}{nh_n}\sum_{l,j}\delta_{lj}K\left(\frac{x_1 - l/n}{h_n}\right)K\left(\frac{x_2 - j/n}{h_n}\right).$$

Following the arguments identical to that in the proof of Lemma 4.4, this expression can be shown to converge to the covariance function of $\bar{W}$. What remains now is the justification of the asymptotic normality of finite dimensional marginals of $\bar{\epsilon}_n$ and proving tightness.

Justifying asymptotic normality of the finite dimensional marginals of $\bar{\epsilon}_n$ requires showing the asymptotic normality of any finite linear combination of marginals of $\bar{\epsilon}_n$ and then applying the Cramér-Wold device. Given the convergence of the covariances, it suffices to prove that for $(c_r, t_r) \in \mathbb{R}, 1 \le r \le R \in \mathbb{N}$,

$$\frac{1}{\sqrt{v_n}}\sum_{r\le R}c_r\bar{\epsilon}_n(t_r) \xrightarrow{d} N(0,1), \tag{C.13}$$

where $v_n^2 =$ Var $\left(\sum_{r\le r}c_r\bar{\epsilon}_n(t_r)\right)$. The left hand side equals $\sum_i w_{in}\epsilon_i$ where

$$w_{in} = \frac{1}{\sqrt{nh_n}v_n}\sum_{r\le R}c_rK\left(\frac{d_0 - i/n}{h_n} + t_r\right).$$

As in (Robinson, 1997, Assumption 2), we assume $\epsilon_i$'s to be a linear process with martingale innovations and square summable coefficients, i.e, there is a sequence of martingale differences $u_j, j \in \mathbb{Z}$ adapted to $\mathcal{F}_j = \sigma\{u_k : k \le j\}$ with mean 0 and variance 1, such that

$$\epsilon_i = \sum_{j=-\infty}^{\infty} \alpha_j u_{i-j}, \quad \sum_{j=-\infty}^{\infty} \alpha_j^2 < \infty. \tag{C.14}$$

To show asymptotic normality, we justify conditions (2.3) and (2.6) from Robinson (1997). The condition (2.3) is just a normalization requirement which holds in our case as the variance of the left hand side of (C.13) is 1. The condition (2.6) of Robinson (1997) is about justifying the existence of a positive-valued sequence $a_n$ such that as $n \to \infty$,

$$\left( \sum_i w_{in}^2 \sum_{|j|>a_n} \alpha_j^2 \right)^{1/2} + \max_{1\le i\le n} |w_{in}| \sum_{|j|\le a_n} |\alpha_j| \to 0. \tag{C.15}$$

For $a_n$ such that $a_n \to \infty$ and $nh_n/a_n \to \infty$, $\sum_{|j|>a_n} \alpha_j^2 = o(1)$, due to (C.14). Also, by Cauchy-Schwartz, $\sum_{|j|\le a} |\alpha_j| = O(\sqrt{a_n})$. By the compactness of the kernel and the fact that $v_n = O(1)$, $\sum_i w_{in}^2 = O(1)$. As the kernel $K$ is bounded, $\max_{1\le i\le n} |w_{in}| = O\left(1/\sqrt{nh_n}\right)$. Hence, the left hand side of (C.15) is $o(1) + O\left(\sqrt{a_n/nh_n}\right)$ which is $o(1)$. This shows convergence of the finite dimensional marginals.

For tightness, recall that for $t \in [-T, T]$

$$\bar{\epsilon}_n(t) = \frac{1}{\sqrt{nh_n}} \sum_{i:|d_0 - i/n| \le (L_0+T)h_n} \epsilon_i K\left(\frac{d_0 - i/n}{h_n} + t\right).$$

We have

$$E\left[\bar{\epsilon}_n(t_1) - \bar{\epsilon}_n(t_2)\right]^2 \quad = \quad \frac{1}{nh_n} \int_{-\pi}^{\pi} \psi(u)(\hat{K}_{x_1}(u) - \hat{K}_{x_2}(u))(\hat{K}_{x_1}(-u) - \hat{K}_{x_2}(-u))du$$

As $\psi$ is a bounded function, the above expression is bounded up to a constant, due to Cauchy-Schwartz, by

$$\frac{1}{nh_n}\int_{-\pi}^{\pi}|\hat{K}_{x_1}(u)-\hat{K}_{x_2}(u)|^2du.$$

As $\hat{K}_{x_1}(u)=\sum_j K\left((x_1-j/n)/h_n\right)e^{\imath ju}$,

$$|\hat{K}_{x_1}(u)-\hat{K}_{x_2}(u)|^2 \lesssim nh_n|t_1-t_2|^{2\alpha}$$

due to Lipschitz continuity of $K$. Hence,

$$E\left[\bar{\epsilon}_n(t_1)-\bar{\epsilon}_n(t_2)\right]^2 \quad\lesssim\quad |t_1-t_2|^{2\alpha}$$

The tightness follows from Theorem 12.3 of Billingsley (1968, pp. 95). $\square$

# APPENDIX D

# Proofs for Chapter 5

## D.1 Proof of Theorem 5.2

Here, we establish consistency with respect to the (stronger) Hausdorff metric,

$$d_H(S_1, S_2) = \max\left[\sup_{x \in S_1} \rho(x, S_2), \sup_{x \in S_2} \rho(x, S_1)\right]. \tag{D.1}$$

Moreover, we would only require $\min(m, n) \to \infty$ instead of taking $m$ to be of the form $m_0 n^\beta$, $\beta > 0$.

To exhibit the dependence on $m$, we will denote $\mathbb{M}_n$ by $\mathbb{M}_{m,n}$. Recall that $M_m(S) = E\left[\mathbb{M}_{m,n}(S)\right]$ converges to $M(S)$ for each $S \in \mathcal{S}$. Also, $\mathrm{Var}(\mathbb{M}_{m,n}(S)) = (1/n)\mathrm{Var}\left((\Phi(\sqrt{m}\bar{Y}_1) - \gamma)1_S(X)\right) \leq 1/n$ which converges to zero. Hence, $\mathbb{M}_{m,n}(S)$ converges in probability to $M(S)$ for any $S \in \mathcal{S}$, as $\min(m, n) \to \infty$.

The space $(\mathcal{S}, d_H)$ is compact (Blaschke Selection theorem) and $M$ is a continuous function on $\mathcal{S}$. The desired result will be a consequence of argmin continuous mapping theorem (van der Vaart and Wellner, 1996, Theorem 3.2.2) provided we can justify that $\sup_{S\in\mathcal{S}}|\mathbb{M}_{m,n}(S)-M(S)|$ converges in probability to zero. To this end, let

$$\mathbb{M}^1_{m,n}(S)=\mathbb{M}_{m,n}(S)+\mathbb{P}_n\gamma 1_X(S)=\mathbb{P}_n\Phi\left(\sqrt{m}\bar{Y}\right)1_S(X)$$

and $\mathbb{M}^1(S)=\mathbb{M}(S)+P\gamma 1_X(S)$. Note that

$$\sup_{S\in\mathcal{S}}|\mathbb{M}_{m,n}(S)-M(S)|\leq\gamma\sup_{S\in\mathcal{S}}|(\mathbb{P}_n-P)(S)|+\sup_{S\in\mathcal{S}}|\mathbb{M}^1_{m,n}(S)-M^1(S)|.$$

The first term in the above expression converges in probability to zero (Ranga Rao, 1962). As for the second term, note that $\mathbb{M}^1_{m,n}(S)$ converges in probability to $M^1(S)$ for each $S$ and $\mathbb{M}^1_{m,n}$ is monotone in $S$, i.e., $\mathbb{M}^1_{m,n}(S_1)\leq\mathbb{M}^1_{m,n}(S_2)$ whenever $S_1\subset S_2$. As the space $(\mathcal{S},d_H)$ is compact, there exist $S(1),\ldots,S(l(\delta))$ such that $\sup_{S\in\mathcal{S}}\min_{1\leq l\leq l(\delta)}d_H(S,S(l))<\delta$, for any $\delta>0$. Hence,

$$\begin{aligned}
&\sup_{S\in\mathcal{S}}|\mathbb{M}^1_{m,n}(S)-M^1(S)|\\
&\quad=\max_{1\leq l\leq l(\delta)}\sup_{d_H(S,S(l))<\delta}|\mathbb{M}^1_{m,n}(S)-M^1(S)|\\
&\quad\leq 2\max_{1\leq l\leq l(\delta)}\sup_{d_H(S,S(l))<\delta}|\mathbb{M}^1_{m,n}(S)-M^1(S(l))|\\
&\quad\leq 2\max_{1\leq l\leq l(\delta)}\max(|\mathbb{M}^1_{m,n}({}_\delta(S(l)))-M^1(S(l))|,|\mathbb{M}^1_{m,n}((S(l))^\delta)-M^1(S(l))|).
\end{aligned}$$

The right side in the above display converges in probability to $2\max_{1\leq l\leq l(\delta)}[\max(|M^1({}_\delta(S(l)))-M^1(S(l))|,|M^1((S(l))^\delta)-M^1(S(l))|)]$ can be made arbitrarily small by choosing small $\delta$ (as $M^1$ is continuous). Also, as the map $S\mapsto d_F(S,S_0)$ from $(\mathcal{S},d_H)$ to $\mathbb{R}$ is continuous, we have consistency in the $d_F$ metric as well. This completes the proof. $\square$

## D.2 Proof of Theorem 5.9

In light of what has been derived in the proof of Theorem 5.2, it suffices to show that $\mathbb{M}_n(S)$ converges in probability to $M(S)$. Note that,

$$E\left[\Phi\left(\sqrt{nh_n^2}\hat{\mu}(x_{kl})\right)\right] = \tilde{\Phi}_n\left(\frac{\sqrt{nh_n^2}\bar{\mu}(x_{kl})}{\sqrt{1+\Sigma_n^2(x_{kl})}}\right).$$

For $x \in \{(x_1, x_2) : k/m \leq x_1 < (k+1)/m, l/m \leq x_2 < (l+1)/m\}$, let

$$\hat{f}_n(x) = \widetilde{W}_{kl} \text{ and } f_n(x) = E\left[\hat{f}_n(x)\right] = \tilde{\Phi}_n\left(\frac{\sqrt{nh_n^2}\bar{\mu}(x_{kl})}{\sqrt{1+\Sigma_n^2(x_{kl})}}\right) - \gamma.$$

Then $\bar{M}_n(S) = \int_S f_n(x)dx$. For any fixed $x$ in the interior of the set $S_0$, $f_n(x) = \tilde{\Phi}_n(0)$ for sufficiently large $n$ which converges to $1/2$. As $\mu$ is continuous, for any fixed $x \notin S_0$, $\mu_{x,\delta_x} = \inf\{\mu(y) : \rho(x,y) < \delta_x\} > 0$ for some $\delta_x > 0$. Hence $f_n(x) \geq \Phi(\sqrt{nh_n^2}\mu_{x,\delta_x})$ converges to 1. Also, $|f_n(x)| \leq 1$ and hence, $\bar{M}_n(S)$ converges to $M(S)$ by DCT.

Moreover,

$$\text{Var}(\mathbb{M}_n(S)) \leq \frac{1}{n^2}\sum_{k,l,k',l'} \text{Cov}\left(\hat{f}_n(x_{k,l}), \hat{f}_n(x_{k',l'})\right).$$

As $|\hat{f}_n(x_{k,l})| \leq 1$, and $\hat{f}_n(x_{k,l})$ and $\hat{f}_n(x_{k',l'})$ are independent whenever $\min\{|k-k'|, |l-l'|\} > 2L_0mh_n$, we have

$$\sum_{k',l'} \text{Cov}\left(\hat{f}_n(x_{k,l}), \hat{f}_n(x_{k',l'})\right) \lesssim (mh_n)^2 = nh_n^2,$$

for any fixed $k$ and $l$. Hence, $\text{Var}(\mathbb{M}_n(S))$is bounded (up to a constant) by $n(nh_n^2)/n^2$ which converges to zero. Hence, $\mathbb{M}_n(S)$ converges in probability to $M(S)$, which completes the proof. $\square$

## D.3 Proof of Lemma 5.11

The sum $\sum_{k,l} W_{kl} 1_S(x_{kl})$ can be written as $\sum_{1\le i,j\le \lceil\sqrt{m'}\rceil} R_{i,j}$ where for $k_i$s and $l_j$s defined as in (5.20), each block

$$R_{i,j} = \sum_{k,l:1\le k_i,l_j\le m} W_{k_i l_j} 1_S(x_{k_i l_j}) \tag{D.2}$$

is a sum of $r_{ij}$ many independent random variables with

$$\left\lfloor m/\left\lceil\sqrt{m'}\right\rceil\right\rfloor^2 \le r_{ij} \le \left\lceil m/\left\lceil\sqrt{m'}\right\rceil\right\rceil^2.$$

As $m/\sqrt{m'} \to \infty$,

$$n/(2m') \le r_{ij} \le 2n/m', \tag{D.3}$$

for large $n$, a fact we use frequently in the proofs. Note that $\sum_{1\le i,j\le\lceil\sqrt{m'}\rceil} r_{ij} = n$ and hence, by convexity of $\exp(\cdot)$,

$$\exp\left(\frac{1}{n}\sum_{1\le k,l\le m} W_{kl}1_S(x_{kl})\right) \le \sum_{1\le i,j\le\lceil\sqrt{m'}\rceil} \frac{r_{ij}}{n}\exp\left(\frac{R_{ij}}{r_{ij}}\right).$$

As $|W_{kl}1_S(x_{kl})| \le 1_S(x_{kl})$,

$$\begin{aligned} P\left(\frac{1}{n}\sum_{k,l} W_{kl}1_S(x_{kl}) \ge a\right) &\le \sum_{1\le i,j\le\lceil\sqrt{m'}\rceil} \frac{r_{ij}}{n} E\ \exp\left(\frac{\lambda R_{ij}}{r_{ij}} - \lambda a\right) \text{ and} \\ E\ \exp\left(\frac{\lambda R_{ij}}{r_{ij}}\right) &\le \exp\left(\frac{\lambda^2}{8r_{ij}^2}\sum_{k,l:1\le k_i,l_j\le m}\left(1_S(x_{k_i l_j})\right)^2\right). \end{aligned}$$

The second bound in the above display is simply the one used in proving Hoeffding's inequality for independent sequences (Hoeffding, 1963, equation (4.16)). Consequently,

$$
\begin{aligned}
&P\left(\frac{1}{n}|\sum_{k,l} W_{kl}1_S(x_{kl}) \geq a\right) \\
&\leq \quad e^{-\lambda a}\exp\left(\frac{\lambda^2}{8(n/2m')^2}\max_{1\leq i,j\leq\lceil\sqrt{m'}\rceil}\left[\sum_{k,l:1\leq k_i,l_j\leq m} 1_S(x_{k_il_j})\right]\right).
\end{aligned}
$$

Choosing

$$
\lambda = \frac{a(n/m')}{\max_{1\leq i,j\leq\lceil\sqrt{m'}\rceil}\left[(1/r_{ij})\sum_{k,l:1\leq k_i,l_j\leq m} 1_S(x_{k_il_j})\right]}
$$

and paralleling the above steps to bound $P\Big((1/n)\sum_{k,l} W_{kl}1_S(x_{kl}) \leq a\Big)$, we get

$$
\begin{aligned}
&P\left(\frac{1}{n}\left|\sum_{k,l} W_{kl}1_S(x_{kl})\right| \geq a\right) \\
&\leq \quad 2\exp\left[-\frac{na^2}{16m'\max_{1\leq i,j\leq\lceil\sqrt{m'}\rceil}\left[(1/r_{ij})\sum_{k,l:1\leq k_i,l_j\leq m} 1_S(x_{k_il_j})\right]}\right].
\end{aligned}
$$

Using the definition of $\tilde{d}_n$, the result follows.

## D.4 Proof of Theorem 5.12

Let $S = \min\{s \geq 1 : 2^{-s}R \leq \delta_2/2\}$. By means of condition (5.22), we can choose $C$ to be a constant large enough so that

$$
\sqrt{n/m'}(\delta_1 - \delta_2) \geq 48\sum_{s=1}^{S} 2^{-s}RH^{1/2}(2^{-s}R, \mathcal{G}, \tilde{d}_n) \vee (1152\log 2)^{1/2}(4m')R.
$$

We denote the class of functions $\mathcal{G}$ by $\{g_\theta : \theta \in \Theta\}$ for convenience. Let $\{g_j^s\}_{j=1}^{N_s}$ be a minimal $2^{-s}R$-covering set of $\tilde{\mathcal{S}}$, $s = 0, 1, \ldots$. So, $N_s = N(2^{-s}R, \tilde{\mathcal{S}}, \tilde{d}_n)$. For, any $\theta \in \Theta$, let $g_\theta^s$ denote approximation of $g_\theta$ from the collection $\{g_j^s\}_{j=1}^{N_s}$. As $|W_{kl}| \leq 1$, applying Cauchy-Schwartz to each block $R_{i,j}$ defined in (D.2) and using (D.3) yields

$$\left| \frac{1}{n} \sum_{k,l} V_{kl}(g_\theta(x_{kl}) - g_\theta^S(x_{kl})) \right| \leq 2\tilde{d}_n(g_\theta, g_\theta^S) \leq \delta_2.$$

Hence, it suffices to prove the exponential inequality for

$$P\left( \max_{j=1,\dots,N_s} \left| \frac{1}{n} \sum_{k,l} V_{kl} g_j^S(x_{kl}) \right| \geq \delta_1 - \delta_2 \right).$$

Next, we use a chaining argument. Note that $g_\theta^S = \sum_{s=1}^S (g_\theta^s - g_\theta^{s-1})$. By triangle inequality,

$$\tilde{d}_n(g_\theta^s, g_\theta^{s-1}) \leq \tilde{d}_n(g_\theta^s, g_\theta) + \tilde{d}_n(g_\theta, g_\theta^{s-1}) \leq 3(2^{-s}R).$$

Let $\eta_s$ be positive numbers satisfying $\sum_{s\leq S} \eta_s \leq 1$. Then,

$$\begin{aligned}
&P^*\left( \sup_{\theta\in\Theta} \left| \frac{1}{n} \sum_{s=1}^S \sum_{k,l} V_{kl}(g_\theta^s(x_{kl}) - g_\theta^{s-1}(x_{kl})) \right| \geq \delta_1 - \delta_2 \right) \\
&\leq \sum_{s=1}^S P^*\left( \sup_{\theta\in\Theta} \left| \frac{1}{n} \sum_{k,l} V_{kl}(g_\theta^s(x_{kl}) - g_\theta^{s-1}(x_{kl})) \right| \geq (\delta_1 - \delta_2)\eta_s \right) \\
&\leq \sum_{s=1}^S 2\exp\left[ 2H(2^{-s}R, \tilde{\mathcal{S}}, \tilde{d}_n) - \frac{n(\delta_1-\delta_2)^2\eta_n^2}{9(16m')2^{-2s}R^2} \right]. \qquad \text{(D.4)}
\end{aligned}$$

We choose $\eta_s$ to be

$$\eta_s = \frac{6\sqrt{16m'}2^{-s}RH^{1/2}(2^{-s}R, \tilde{\mathcal{S}}, \tilde{d}_n)}{\sqrt{n}(\delta_1 - \delta_2)} \vee \frac{2^{-s}\sqrt{s}}{8}.$$

The rest of the argument is identical to that Lemma 3.2 of van de Geer (2000). It can be shown that $\sum_{s\leq S} \eta_s \leq 1$. Moreover, the above choice of $\eta_s$ guarantees

$$H(2^{-s}R, \tilde{\mathcal{S}}, \tilde{d}_n) \leq \frac{n(\delta_1-\delta_2)^2\eta_s^2}{36(16m'^2)2^{-2s}R^2}.$$

Hence the bound in (D.4) is at most

$$\sum_{s=1}^{s} 2 \exp \left[ -\frac{n(\delta_1 - \delta_2)^2 \eta_s^2}{18(16m')2^{-2s} R^2} \right].$$

Next, using $\eta_s \leq 2^{-s}\sqrt{s}/8$ and that $n(\delta_1 - \delta_2^2)/(1152(16m')R^2) \geq \log(2)$, it can be shown that the above display is bounded above by

$$\begin{aligned}
& \sum_{s=1}^{\infty} 2 \exp \left[ -\frac{n(\delta_1 - \delta_2)^2 s}{1152(16m')R^2} \right] \\
\leq \;& 2 \left( 1 - \exp \left[ -\frac{n(\delta_1 - \delta_2)^2}{1152(16m')R^2} \right] \right)^{-1} \exp \left[ -\frac{n(\delta_1 - \delta_2)^2 s}{1152(16m')R^2} \right] \\
\leq \;& 4 \exp \left[ -\frac{n(\delta_1 - \delta_2)^2 s}{1152(16m')R^2} \right].
\end{aligned}$$

This completes the proof.

## D.5 Proof of Proposition 5.16

Note that $\sqrt{mn}(\hat{\tau} - \tau_0) = O_P(1)$. So, given $\alpha > 0$, there exists $L_\alpha > 0$ such that for $V_{n,\alpha} = [\tau_0 - L_\alpha/\sqrt{mn}, \tau_0 + L_\alpha/\sqrt{mn}]$, $P[\hat{\tau} \in V_{n,\alpha}] > 1 - \alpha$. Let $\hat{S}_n(\tau)$ denote the estimate of $S_0$ based on $\mathbb{M}_n(S, \tau)$. Then,

$$P^* \left[ d(\hat{S}_n(\hat{\tau}), S_0) > \delta_n \right] \;\leq\; P^* \left[ d(\hat{S}_n(\hat{\tau}), S_0) > \delta_n, \hat{\tau} \in V_{n,\alpha} \right] + \alpha.$$

Following the arguments for the proof of Proposition 5.6, the outer probability on the right side can be bounded by

$$\sum_{k=0}^{k_n} P^* \left( \inf_{A \in \mathcal{S}_{n,k}} \mathbb{M}_n(A, \hat{\tau}) - \mathbb{M}_n(S_0, \hat{\tau}) \leq 0, \hat{\tau} \in V_{n,\alpha} \right)$$

The is further bounded by:

$$\sum_{k=0}^{k_n} P^* \left( \sup_{\substack{A \in \mathcal{S}_{n,k}, \\ \tau \in V_{n,\alpha}}} |(\mathbb{M}_n(S,\tau) - M(S)) - (\mathbb{M}_n(S_0,\tau) - M(S_0))| > \inf_{A \in \mathcal{S}_{n,k}} (M(S) - M(S_0)) \right). \tag{D.5}$$

As before, $M(S) - M(S_0) \geq c(\gamma) F(S \triangle S_0)$, and hence, (D.5) is bounded by

$$\begin{aligned}
&\sum_{k=0}^{k_n} P^* \left( \sup_{\substack{A \in \mathcal{S}_{n,k}, \\ \tau \in V_{n,\alpha}}} |(\mathbb{M}_n - M_m)(S,\tau) - (\mathbb{M}_n - M_m)(S_0,\tau)| > c(\gamma) 2^k \delta_n / 3 \right) \\
&+ \sum_{k=0}^{k_n} 1 \left[ \sup_{\substack{A \in \mathcal{S}_{n,k}, \\ \tau \in V_{n,\alpha}}} |(M_m(S,\tau) - M_m(S,\tau_0)) - (M_m(S_0,\tau) - M_m(S_0,\tau_0))| \geq c(\gamma) 2^k \delta_n / 3 \right] \\
&+ \sum_{k=0}^{k_n} 1 \left[ \sup_{A \in \mathcal{S}_{n,k}} |(M_m - M)(S,\tau_0) - (M_m - M)(S_0,\tau_0)| \geq c(\gamma) 2^k \delta_n / 3 \right].
\end{aligned} \tag{D.6}$$

The third term can be shown to be zero for sufficiently large $n$ in the same manner as in the proof of Proposition 5.6. Note that the first term can be written as

$$\sum_{k=0}^{k_n} P^* \left( \sup_{A \in \mathcal{S}_{n,k}, \tau \in V_{n,\alpha}} \left| \mathbb{G}_n g_{n,\tau}(\bar{Y}) 1_{A \triangle S_0}(X) \right| > c(\gamma) 2^{k-1} \delta_n \sqrt{n} / 3 \right), \tag{D.7}$$

where $g_{n,\tau}(y) = \Phi\left(\sqrt{m}(y - \tau)\right) - \gamma$. We are now in a position to apply Theorem 5.5 to each term of (D.7). In the setup of Theorem 5.5, $N = 2^{k-1} \delta_n \sqrt{n}$ and the concerned class of functions is $\mathcal{G}_{n,k} = \{g_{n,\tau}(\bar{Y}) 1_B(X) : B = A \triangle S_0, A \in \mathcal{S}_{n,k}, \tau \in V_{n,\alpha}\}$. For $B \in \{A \triangle S_0 : A \in \mathcal{S}_{n,k}\}$, $\|g_{n,\tau} 1_B\|_{L_2(P)} \leq [E 1_B(X)]^{1/2} \leq (2^{k+1} \delta_n)^{1/2}$. So, we can choose $R = R_{n,k} = (2^{k+1} \delta_n)^{1/2}$. Also,

$$H_B(u, \{A \triangle S_0 : A \in \mathcal{S}_{n,k}\}, L_2(P)) \leq A_0 u^{-1},$$

for some constant $A_0 > 0$. To bound the entropy of the class of functions $\mathcal{T}_n = \{g_{n,\tau}(\cdot) : \tau \in V_{n,\alpha}\}$, let $\tau_0 - L_\alpha / \sqrt{mn} = t_0 < t_1 < \ldots < t_{r_n} = \tau_0 + L_\alpha / \sqrt{mn}$ be such that $|t_i - t_{i-1}| \leq u/\sqrt{m}$, for $u > 0$. Note that $r_n \leq 4L_\alpha u/\sqrt{n}$. As $\Phi$ is Lipschitz continuous of order 1 (with Lipschitz constant bounded by 1),

$$|g_{n,\tau}(\bar{y}) - g_{n,\tau_i}(\bar{y})| \leq \sqrt{m}|\tau - \tau_i| \leq u,$$

for $\tau \in [\tau_i, \tau_{i+1}]$. Hence,

$$H_B(u, \mathcal{T}_n, L_2(P)) \leq A_1 \log(u/\sqrt{n}),$$

for some constant $A_1 > 0$ and for small $u > 0$. As the class $\mathcal{G}_{n,k}$ is formed by product of the two classes $\mathcal{S}_{n,k}$ and $\mathcal{T}_n$, the bracketing number for $\mathcal{G}_{n,k}$ is bounded above by,

$$H_B(u, \mathcal{G}_{n,k}, L_2(P)) \leq A_0 u^{-1} + A_1 \log(u/\sqrt{n}) \leq A_2 u^{-1}.$$

In light of the above bound on the entropy, the first term in (D.6) can be shown to go to zero by arguing in the same manner as in the proof of Proposition 5.6.

For the second term in (D.6), note that $|\Phi\left(\sqrt{m}(\bar{Y} - \tau)\right) - \Phi\left(\sqrt{m}(\bar{Y} - \tau_0)\right)| \leq \sqrt{m}|\tau_0 - \tau|$. Hence,

$$\begin{aligned}
&|(M_m(S,\tau) - M_m(S,\tau_0)) - (M_m(S_0,\tau) - M_m(S_0,\tau_0))| \\
&= \left|P_m\left[\left\{\Phi\left(\sqrt{m}(\bar{Y} - \tau)\right) - \Phi\left(\sqrt{m}(\bar{Y} - \tau_0)\right)\right\}\{1_S(X) - 1_{S_0}(X)\}\right]\right| \\
&\leq \sqrt{m}|\tau_0 - \tau|\,|P_m\,|1_S(X) - 1_{S_0}(X)||\,.
\end{aligned}$$

Thus, the second term in (D.6) is bounded by

$$2(L_\alpha/\sqrt{n}) \sup_{S \in \mathcal{S}_{n,k}} P_m|1_S(X) - 1_{S_0}(X)| \leq L_\alpha 2^{k+2}\delta_n/\sqrt{n}.$$

This is eventually smaller that $c(\gamma)2^k\delta_n/3$ and hence, each term in the second sum of (D.6) is eventually zero. As $\alpha$ is arbitrary, we get the result. $\square$

## D.6 Proof of Proposition 5.17

Note that $\sqrt{n}(\hat{\tau}-\tau_0) = O_P(1)$. So, given $\alpha > 0$, there exists $L_\alpha > 0$ such that for $V_{n,\alpha} = [\tau_0 - L_\alpha/\sqrt{n}, \tau_0 + L_\alpha/\sqrt{n}]$, $P[\hat{\tau} \in V_{n,\alpha}] > 1-\alpha$. Let $\hat{S}_n(\tau)$ denote the estimate of $S_0$ based on $\mathbb{M}_n(S,\tau)$. We have,

$$P^*\left[d(\hat{S}_n(\hat{\tau}), S_0) > \delta_n\right] \;\; \leq \;\; P^*\left[\delta_n < d(\hat{S}_n(\hat{\tau}), \hat{\tau} \in V_{n,\alpha}\right] + \alpha.$$

Following the arguments for the proof of Proposition 5.13, the first term can be bounded by

$$\sum_{k\geq 0, 2^k\delta_n \leq 1} P^*\left(\inf_{A\in\mathcal{S}_{n,k}} \mathbb{M}_n(A,\hat{\tau}) - \mathbb{M}_n(S_0,\hat{\tau}) \leq 0, \hat{\tau}\in V_{n,\alpha}\right)$$

This is at most

$$\sum_{k=0}^{k_n} P^*\left(\sup_{\substack{A\in\mathcal{S}_{n,k},\\ \tau\in V_{n,\alpha}}} |(\mathbb{M}_n(S,\tau)-M(S)) - (\mathbb{M}_n(S_0,\tau)-M(S_0))| > \inf_{A\in\mathcal{S}_{n,k}} (M(S)-M(S_0))\right). \tag{D.8}$$

Note that $M(S)-M(S_0) \geq c(\gamma)\lambda(S\triangle S_0)$ as earlier, and hence (D.8) is bounded by

$$\sum_{k=0}^{k_n} P^*\left(\sup_{\substack{A\in\mathcal{S}_{n,k},\\ \tau\in V_{n,\alpha}}} |(\mathbb{M}_n(S,\tau)-M(S)) - (\mathbb{M}_n(S_0,\tau)-M(S_0))| > c(\gamma)2^k\delta_n\right).$$

Moreover,

$$\begin{aligned}
&|(\mathbb{M}_n(S,\tau)-M(S)) - (\mathbb{M}_n(S_0,\tau)-M(S_0))| \\
&\qquad\qquad \leq \;\; |(\mathbb{M}_n(S,\tau_0)-M(S)) - (\mathbb{M}_n(S_0,\tau_0)-M(S_0))| \\
&\qquad\qquad\qquad + \left|(\mathbb{M}_n(S,\tau)-\mathbb{M}(S,\tau_0)) - (\mathbb{M}_n(S_0,\tau)-\mathbb{M}_(S_0,\tau_0))\right|.
\end{aligned}$$

By the Lipschitz continuity of $\Phi$, we have

$$\begin{aligned}
&\big|(\mathbb{M}_n(S,\tau) - \mathbb{M}(S,\tau_0)) - (\mathbb{M}_n(S_0,\tau) - \mathbb{M}_(S_0,\tau_0))\big| \\
&\qquad\qquad \leq \quad \sqrt{nh_n^2}|\tau - \tau_0| \left[\frac{1}{n}\sum_{k,l}|1_S(x_{kl}) - 1_{S_0}(x_{kl})|\right] \\
&\qquad\qquad \leq \quad \sqrt{nh_n^2}|\tau - \tau_0| \left[\lambda(S\triangle S_0) + O\left(\frac{1}{\sqrt{n}}\right)\right].
\end{aligned}$$

Here, the last step follows from calculations similar to those in the proof of Lemma 5.10. Consequently, for sufficiently large $n$,

$$\begin{aligned}
&\sup_{\substack{A\in\mathcal{S}_{n,k},\\ \tau\in V_{n,\alpha}}} \big|(\mathbb{M}_n(S,\tau) - \mathbb{M}(S,\tau_0)) - (\mathbb{M}_n(S_0,\tau) - \mathbb{M}_(S_0,\tau_0))\big| \\
&\qquad\qquad \leq \quad (2L_\alpha)h_n\left[2^{k+1}\delta_n + O\left(\frac{1}{\sqrt{n}}\right)\right] < \frac{c(\gamma)}{2}2^k\delta_n.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sum_{k=0}^{k_n} P^*\left(\sup_{\substack{A\in\mathcal{S}_{n,k},\\ \tau\in V_{n,\alpha}}} |(\mathbb{M}_n(S,\tau) - M(S)) - (\mathbb{M}_n(S_0,\tau) - M(S_0))| > c(\gamma)2^k\delta_n\right) \\
&\quad \leq \quad \sum_{k=0}^{k_n} P^*\left(\sup_{\substack{A\in\mathcal{S}_{n,k},\\ \tau\in V_{n,\alpha}}} |(\mathbb{M}_n(S,\tau_0) - M(S)) - (\mathbb{M}_n(S_0,\tau_0) - M(S_0))| > \frac{c(\gamma)}{2}2^k\delta_n\right).
\end{aligned}$$

The above is a probability inequality based on the criterion with known $\tau_0$. This can be shown to go to zero by calculations identical to those in the proof of Proposition 5.13. As $\alpha > 0$ is arbitrary, we get the result. $\square$

# APPENDIX E

# Proofs for Chapter 6

## E.1 Proof of Theorem 6.1

Note that if $\kappa_n r_n = O(1)$, i.e., there exists $C > 0$, such that $\kappa_n r_n \leq C$ for all $n$, then

$$\begin{aligned} P\left(r_n \rho_n(\hat{d}_n, d_n) \geq C\right) &= P\left(r_n \kappa_n \rho_n(\hat{d}_n, d_n) \geq C\kappa_n\right) \\ &\leq P\left(\rho_n(\hat{d}_n, d_n) \geq \kappa_n\right), \end{aligned}$$

which converges to zero. Therefore, the conclusion of the theorem is immediate when $\kappa_n r_n = O(1)$. Hence, we only need to address the situation where $\kappa_n r_n \to \infty$.

For a fixed realization of $\hat{\theta} = \theta$, we use $\hat{d}_n(\theta)$ to denote our estimate, so that $\hat{d}_n = \hat{d}_n(\hat{\theta}_n)$. For any $L > 0$,

$$\begin{aligned} P\left(r_n \rho_n(\hat{d}_n(\hat{\theta}_n), d_n) \geq 2^L\right) &\leq P\left(r_n \kappa_n > r_n \rho_n(\hat{d}_n(\hat{\theta}_n), d_n) \geq 2^L, \hat{\theta}_n \in \Theta_n^\tau\right) \\ &\quad + P\left(\rho_n(\hat{d}_n(\hat{\theta}_n), d_n) \geq \kappa_n\right) + \tau. \end{aligned} \tag{E.1}$$

The second term on the right side goes to zero. Further,

$$\begin{aligned}
&P\left(r_n\kappa_n > r_n\rho_n(\hat{d}_n(\hat{\theta}_n), d_n) \geq 2^L, \hat{\theta}_n \in \Theta_n^\tau\right) \\
&\qquad = E\left[P\left(r_n\kappa_n > r_n\rho_n(\hat{d}_n(\hat{\theta}_n), d_n) \geq 2^L \mid \hat{\theta}_n\right) 1\left[\hat{\theta}_n \in \Theta_n^\tau\right]\right] \\
&\qquad \leq \sup_{\theta\in\Theta_n^\tau} P\left(r_n\kappa_n > r_n\rho_n(\hat{d}_n(\theta), d_n) \geq 2^L\right). \qquad \text{(E.2)}
\end{aligned}$$

Let $S_{j,n} = \{d : 2^j \leq r_n\rho_n(d, d_n) < \min(2^{j+1}, \kappa_n r_n)\}$ for $j \in \mathbb{Z}$. If $r_n\rho_n(\hat{d}_n(\theta), d_n)$ is larger than $2^L$ for a given positive integer $L$ (and smaller than $\kappa_n r_n$), then $\hat{d}_n(\hat{\theta}_n)$ is in one of the shells $S_{j,n}$'s for $j \geq L$. By definition of $\hat{d}_n(\theta)$, the infimum of the map $d \mapsto \mathbb{M}_n(d,\theta) - \mathbb{M}_n(d_n,\theta)$ over the shell containing $\hat{d}_n(\theta)$ (intersected with $\mathcal{D}_\theta$) is not positive. For $\theta \in \Theta_n^\tau$,

$$\begin{aligned}
&P\left(r_n\kappa_n > r_n\rho_n(\hat{d}_n(\theta), d_n) \geq 2^L\right) \\
&\quad \leq \sum_{j\geq L, 2^j\leq\kappa_n r_n} P^*\left(\inf_{d\in S_{j,n}\cap\mathcal{D}_\theta} \mathbb{M}_n(d,\theta) - \mathbb{M}_n(d_n,\theta) \leq 0\right).
\end{aligned}$$

For every $j$ involved in the sum, $n > N_\tau$ and any $\theta \in \Theta_n^\tau$, (6.2) gives

$$\inf_{2^j/r_n\leq\rho_n(d,d_n)<\min(2^{j+1},\kappa_n r_n)/r_n, d\in\mathcal{D}_\theta} M_n(d,\theta) - M_n(d_n,\theta) \geq c_\tau \frac{2^{2j}}{r_n^2}. \qquad \text{(E.3)}$$

Also, for such a $j$, $n > N_\tau$ and $\theta \in \Theta_n^\tau$,

$$\begin{aligned}
&P^*\left(\inf_{d\in S_{j,n}\cap\mathcal{D}_\theta} \mathbb{M}_n(d,\theta) - \mathbb{M}_n(d_n,\theta) \leq 0\right) \\
&\quad \leq P^*\left(\inf_{d\in S_{j,n}\cap\mathcal{D}_\theta} [(\mathbb{M}_n(d,\theta) - M_n(d,\theta)) - (\mathbb{M}_n(d_n,\theta) - M_n(d_n,\theta))]\right. \\
&\qquad \left. \leq -\inf_{d\in S_{j,n}\cap\mathcal{D}_\theta} M_n(d,\theta) - M_n(d_n,\theta)\right) \\
&\quad \leq P^*\left(\inf_{d\in S_{j,n}\cap\mathcal{D}_\theta} [(\mathbb{M}_n(d,\theta) - M_n(d,\theta)) - (\mathbb{M}_n(d_n,\theta) - M_n(d_n,\theta))] \leq -c_\tau \frac{2^{2j}}{r_n^2}\right) \\
&\quad \leq P^*\left(\sup_{d\in S_{j,n}\cap\mathcal{D}_\theta} |(\mathbb{M}_n(d,\theta) - M_n(d,\theta)) - (\mathbb{M}_n(d_n,\theta) - M_n(d_n,\theta))| \geq c_\tau \frac{2^{2j}}{r_n^2}\right).
\end{aligned}$$

For $n > N_\tau$, by Markov inequality and (6.3), we get

$$\begin{aligned}\sup_{\theta \in \Theta_n^\tau} \sum_{j \geq L, 2^j \leq \kappa_n r_n} & P^* \left( \inf_{d \in S_{j,n} \cap \mathcal{D}_\theta} \mathbb{M}_n(d, \theta) - \mathbb{M}_n(d_n, \theta) \leq 0 \right) \\ & \leq \;\; C_\tau \sum_{j \geq L, 2^j \leq \kappa_n r_n} \frac{\phi_n(\min(2^{j+1}, r_n \kappa_n)/r_n) r_n^2}{c_\tau \sqrt{n} 2^{2j}}. \end{aligned} \tag{E.4}$$

Note that $\phi_n(c\delta) \leq c^\alpha \phi_n(\delta)$ for every $c > 1$. As $\kappa_n r_n \to \infty$, there exists $\bar{N} \in \mathbb{N}$, such that $\kappa_n r_n > 1$. Hence, for $L > 0$ and $n > \max(\bar{N}, N_\tau)$, the above display is bounded by

$$\frac{C_\tau}{c_\tau} \sum_{j \geq L, 2^j \leq \kappa_n r_n} \min(2^{j+1}, r_n \kappa_n) 2^{-2j} \leq \frac{C_\tau}{c_\tau} \sum_{j \geq L, 2^j \leq \kappa_n r_n} 2^{(j+1)\alpha - 2j},$$

by the definition of $r_n$. For any fixed $\eta > 0$, take $\tau = \eta/3$ and choose $L_\eta > 0$ such that the sum on the right side is less than $\eta/3$. Also, there exists $\tilde{N}_\eta \in \mathbb{N}$ such that for all $n > \tilde{N}_\eta \in \mathbb{N}$,

$$P\left( \rho_n(\hat{d}_n(\hat{\theta}_n), d_n) \geq \kappa_n \right) < \eta/3.$$

Hence, for $n > \max(\bar{N}, N_{\eta/3}, \tilde{N}_\eta)$,

$$P\left( r_n \rho_n(\hat{d}_n(\hat{\theta}_n), d_n) > 2^{L_\eta} \right) < \eta,$$

by (E.1) and (E.4). Thus, we get the result when conditions (6.2) and (6.3) hold for some sequence $\kappa_n > 0$.

Further, note that if the conditions in part (b) of the theorem hold for all sequences $\kappa_n > 0$, following the arguments in (E.1) and (E.2), we have

$$P\left( r_n \rho_n(\hat{d}_n(\hat{\theta}_n), d_n) > 2^L \right) \;\; \leq \;\; \sup_{\theta \in \Theta_n^\tau} P\left( r_n \rho_n(\hat{d}_n(\theta), d_n) > 2^L \right) + \tau.$$

Moreover, the bounds in (E.3) and (E.4) hold for all $j \geq L$ and $n > N_\tau$. Hence, we do not need address the event $P\left( \rho_n(\hat{d}_n(\hat{\theta}_n), d_n) \geq \kappa_n \right)$ in (E.1) separately and thus, the result follows. □

## E.2 Proof of Lemma 6.2

Note that $\mathbb{M}_n(\hat{d}_n(\hat{\theta}_n), \hat{\theta}_n) - \mathbb{M}_n(d_n, \hat{\theta}_n)$ is not positive by definition of $\hat{d}_n(\hat{\theta}_n)$. Hence,

$$\begin{aligned}
&P\left[\rho_n(\hat{d}_n(\hat{\theta}_n), d_n) \geq \kappa_n, \hat{\theta}_n \in \Theta_n^\tau\right] \\
&\leq \; E\left[P\left[\rho_n(\hat{d}_n(\hat{\theta}_n), d_n) \geq \kappa_n \mid \hat{\theta}_n\right] 1\left[\hat{\theta}_n \in \Theta_n^\tau\right]\right] \\
&\leq \; \sup_{\theta \in \Theta_n^\tau} P\left[2\rho_n(\hat{d}_n(\theta), d_n) \geq \kappa_n\right] \\
&\leq \; \sup_{\theta \in \Theta_n^\tau} P\left[M_n(\hat{d}_n(\theta), \theta) - M_n(d_n, \theta) \geq c_n^\tau(\kappa_n)\right] \\
&\leq \; \sup_{\theta \in \Theta_n^\tau} P\left[M_n(\hat{d}_n(\theta), \theta) - M_n(d_n, \theta) - \left(\mathbb{M}_n(\hat{d}_n(\theta), \theta) - \mathbb{M}_n(d_n, \theta)\right) \geq c_n^\tau(\kappa_n)\right] \\
&\leq \; \sup_{\theta \in \Theta_n^\tau} P\left[2 \sup_{d \in \mathcal{D}_\theta} |\mathbb{M}_n(d, \theta) - M_n(d, \theta)| \geq c_n^\tau(\kappa_n)\right].
\end{aligned}$$

As the probability in right side converges to zero and $\tau > 0$ is arbitrary, we get the result. □

## E.3 Proof of Theorem 6.3

As sum of tight processes is tight, it suffices to show tightness of $\zeta_n(\cdot, \hat{\theta}_n)$ and $\mathbb{G}_n f_{n,\cdot,\hat{\theta}_n}$ separately. To justify tightness of the process $\zeta_n$, we need to bound

$$P^*\left[\sup_{\tilde{\rho}(h_1,h_2)<\delta_n} \left|\zeta_n(h_1, \hat{\theta}_n) - \zeta_n(h_2, \hat{\theta}_n)\right| > t\right],$$

for $\delta_n \downarrow 0$ and $t > 0$. The above display is bounded by

$$P^*\left[\sup_{\tilde{\rho}(h_1,h_2)<\delta_n} \left|\zeta_n(h_1, \hat{\theta}_n) - \zeta_n(h_2, \hat{\theta}_n)\right| > t, \hat{\theta}_n \in \Theta_n^\tau\right] + P[\hat{\theta}_n \notin \Theta_n^\tau]$$

$$\leq \ 1\left[\sup_{\substack{\theta\in\Theta_n^\tau\\ \tilde\rho(h_1,h_2)<\delta_n}} |\zeta_n(h_1,\theta)-\zeta_n(h_2,\theta)|>t\right]+\tau.$$

By (6.10), the above can be made arbitrarily small for large $n$ and hence, the process $\zeta_n(\cdot,\hat\theta_n)$ is asymptotically tight.

We justify tightness of the process $\{\mathbb{G}_n f_{n,h,\hat\theta} : h\in\mathcal{H}\}$ when (6.11) holds. The proof under the condition on bracketing numbers follows along similar lines. Consider the expression

$$P^*\left[\sup_{\tilde\rho(h_1,h_2)<\delta_n}\left|\mathbb{G}_n(f_{n,h_1,\hat\theta_n}-f_{n,h_2,\hat\theta_n})\right|>t\right],$$

for $\delta_n\downarrow 0$ and $t>0$. Let $e_i, i\geq 1$ denote Rademacher random variables independent of $V$'s and $\hat\theta$. Following the arguments for the proof of the symmetrization lemma for probabilities, Lemma 2.3.7 of van der Vaart and Wellner (1996), for sufficiently large $n$, the above display can be bounded by

$$4P^*\left[\sup_{\tilde\rho(h_1,h_2)<\delta_n}\left|\frac{1}{\sqrt n}\sum_{i=1}^n e_i(f_{n,h_1,\hat\theta}(V_i)-f_{n,h_2,\hat\theta}(V_i))\right|>\frac{t}{4}\right]. \tag{E.5}$$

The only difference from the proof of the cited lemma is that the arguments are to be carried out for fixed realizations of $V_i$'s and $\hat\theta$ (and then outer expectations are taken) instead of just $V_i$'s. Further, from the measurability assumption, the map

$$(V_1,V_2,\ldots,V_n,\hat\theta,e_1,\ldots,e_n)\mapsto \sup_{\tilde\rho(h_1,h_2)<\delta_n}\left|\frac{1}{\sqrt n}\sum_{i=1}^n e_i(f_{n,h_1,\hat\theta}(V_i)-f_{n,h_2,\hat\theta}(V_i))\right|$$

is jointly measurable. Hence, the expression in (E.5) is a probability. Let $Q_n$ denote the marginal distribution of $\hat\theta_n$. Then,

$$\begin{aligned}
&4P\left[\sup_{\tilde\rho(h_1,h_2)<\delta_n}\left|\frac{1}{\sqrt n}\sum_{i=1}^n e_i(f_{n,h_1,\hat\theta}(V_i)-f_{n,h_2,\hat\theta}(V_i))\right|>\frac{t}{4}\right]\\
&= \ 4\int P\left[\sup_{\tilde\rho(h_1,h_2)<\delta_n}\left|\frac{1}{\sqrt n}\sum_{i=1}^n e_i(f_{n,h_1,\theta}(V_i)-f_{n,h_2,\theta}(V_i))\right|>\frac{t}{4}\right]Q_n(d\theta)
\end{aligned}$$

$$\leq \ 4 \sup_{\theta\in\Theta_n^\tau} P\left[\sup_{\tilde{\rho}(h_1,h_2)<\delta_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i(f_{n,h_1,\theta}(V_i)-f_{n,h_2,\theta}(V_i))\right|>\frac{t}{4}\right]+\tau$$

For a fixed $\theta \in \Theta_n^\tau$, let $\mathcal{F}_{n,\theta,\delta_n} = \{f_{n,h_1,\theta} - f_{n,h_2,\theta} : \tilde{\rho}(h_1,h_2) < \delta_n\}$. For $g \in \mathcal{F}_{n,\theta,\delta_n}$, the sum $\sum_{i=1}^n e_i g(V_i)$ (given $V_i$s) is sub-Gaussian and hence, by chaining, Corollary 2.2.8 of van der Vaart and Wellner (1996), the above display can be bounded by

$$\frac{16}{t}\sup_{\theta\in\Theta_n^\tau} E\int_0^{\xi_n(\theta)}\sqrt{\log N\left(u,\mathcal{F}_{n,\theta,\delta_n},L_2(\mathbb{P}_n)\right)}du, \tag{E.6}$$

with

$$\xi_n^2(\theta) = \sup_{g\in\mathcal{F}_{n,\theta,\delta_n}}\|g\|^2_{L_2(\mathbb{P}_n)} = \sup_{g\in\mathcal{F}_{n,\theta,\delta_n}}\left[\frac{1}{n}\sum_{i=1}^n g^2(V_i)\right].$$

The integrand in (E.6) can be bounded using the inequality $N(u,\mathcal{F}^2_{n,\theta,\delta_n},L_2(\mathbb{P}_n)) \leq N^2(u/2,\mathcal{F}_{n,\theta},L_2(\mathbb{P}_n))$. By a change of variable, the expression in (E.6) is then bounded by:

$$\frac{16}{t}\sup_{\theta\in\Theta_n^\tau} E\left[\|F_{n,\theta}\|_{L_2(\mathbb{P}_n)}\int_0^{\xi_n(\theta)/\|F_{n,\theta}\|_{L_2(\mathbb{P}_n)}}\sup_Q\sqrt{\log N\left(u\|F_{n,\theta}\|_{L_2(Q)},\mathcal{F}^2_{n,\theta,\delta_n},L_2(Q)\right)}du\right]. \tag{E.7}$$

By Cauchy-Schwartz inequality, the expectation above is bounded by the product of $E\left[\|F_{n,\theta}\|^2_{L_2(\mathbb{P}_n)}\right]^{1/2}$ and

$$\left[E\left[\int_0^{\xi_n(\theta)/\|F_{n,\theta}\|_{L_2(\mathbb{P}_n)}}\sup_Q\sqrt{\log N\left(u\|F_{n,\theta}\|_{L_2(Q)},\mathcal{F}^2_{n,\theta,\delta_n},L_2(Q)\right)}du\right]^2\right]^{1/2}.$$

By dominated convergence, it can then be shown that the expression in (E.6) goes to zero provided $\sup_{\theta\in\Theta_n^\tau} E\|F_{n,\theta}\|^2_{L_2(\mathbb{P}_n)}$ is O(1) and $\sup_{\theta\in\Theta_n^\tau} P^*\left[\xi_n(\theta)/\|F_{n,\theta}\|_{L_2(\mathbb{P}_n)} > \eta\right]$ goes to zero for any $\eta > 0$. Note that $E\|F_{n,\theta}\|^2_{L_2(\mathbb{P}_n)} = PF^2_{n,\theta}$ which is uniformly bounded in $\theta$ using (6.7). Moreover, the envelopes $F_{n,\theta}$ can be chosen bounded away from zero without disturbing the assumptions of the theorem ($F_{n,\theta} \vee 1$ is also an envelope). Hence, it suffices to show that $\sup_{\theta\in\Theta_n^\tau} E^*\xi_n(\theta)^2$ converges to zero. Note that

$$E^*\xi_n(\theta)^2 \leq E^* \sup_{g\in\mathcal{F}_{n,\theta,\delta_n}} |(\mathbb{P}_n - P)g^2| + \sup_{g\in\mathcal{F}_{n,\theta,\delta_n}} |Pg^2|$$

By (6.9), the second term on the right side goes to zero uniformly in $\theta \in \Theta_n^\tau$. By the symmetrization lemma for expectations, Lemma 2.3.1 of van der Vaart and Wellner (1996), the first term on the right side is bounded by

$$2E^* \sup_{g\in\mathcal{F}^2_{n,\theta,\delta_n}} \left|\frac{1}{n}\sum_{i=1}^n e_i g(V_i)\right|$$

Note that $G_{n,\theta} = (2F_{n,\theta})^2$ is an envelope for the class $\mathcal{F}^2_{n,\theta,\delta_n}$. By condition (6.8), there exists a sequence of numbers $\eta_n \downarrow 0$ (slowly enough) such that $\sup_{\theta\in\Theta_n^\tau} PF^2_{n,\theta}1\left[F^2_{n,\theta} > \eta_n\sqrt{n}\right]$ converges to zero. Let $\mathcal{F}^2_{n,\theta,\delta_n,\eta_n} = \left\{g1[g \leq n\eta_n^2] : \; g \in \mathcal{F}^2_{n,\theta,\delta_n}\right\}$. Then, the above display is bounded by:

$$2E^* \sup_{g\in\mathcal{F}^2_{n,\theta,\delta_n,\eta_n}} \left|\frac{1}{n}\sum_{i=1}^n e_i g(V_i)\right| + 4P^*G_{n,\theta}1\left[G_{n,\theta} > n\eta_n^2\right]$$

The second term in the above display goes to zero (uniformly in $\theta$) by (6.8). By the $P$-measurability of the class $\mathcal{F}^2_{n,\theta,\delta_n,\eta_n}$, the first term in the above display is an expectation. For $u > 0$, let $\mathcal{G}_u$ be a minimal $u$-net in $L_1(\mathbb{P}_n)$ over $\mathcal{F}^2_{n,\theta,\delta_n,\eta_n}$. Note that cardinality of $\mathcal{G}_u$ is $N(u, \mathcal{F}^2_{n,\theta,\delta_n,\eta_n}, L_1(\mathbb{P}_n))$ and that

$$2E^* \sup_{g\in\mathcal{F}^2_{n,\theta,\delta_n,\eta_n}} \left|\frac{1}{n}\sum_{i=1}^n e_i g(V_i)\right| \;\; \leq \;\; \frac{2}{n}E^* \sup_{g\in\mathcal{G}_u} \left|\sum_{i=1}^n e_i g(V_i)\right| + u.$$

Also, $\sup_{g\in\mathcal{F}^2_{n,\theta,\delta_n,\eta_n}} |g| \leq n\eta_n^2$. Now, by arguments similar to those in the proof of Theorem 2.4.3 of van der Vaart and Wellner (1996), the first term on the right side is bounded (up to a constant multiple) by

$$E\frac{\eta_n^2 n}{n}\sqrt{1+\log N(u,\mathcal{F}^2_{n,\theta,\delta_n,\eta_n},L_1(\mathbb{P}_n))} \;\leq\; \eta_n^2\sqrt{1+2\sup_Q \log N(u,\mathcal{F}_{n,\theta},L_2(Q))}.$$

As $\eta_n$ converges to zero (uniformly in $\theta$) and $u$ is arbitrary, we get the result. $\square$

## E.4 Proof of Lemma 6.4

Here, we prove a more general result which also applies to situations where the limit distribution of the first stage estimate can appear in the limit process. We state the result below.

**Lemma E.1.** *For a generic $\theta$, let $\Delta_\theta = n^\nu(\theta - \theta_n)$. Consider the setup of Theorem 6.3. Additionally, assume that*

1. *$\Delta_{\hat{\theta}_n} = n^\nu(\hat{\theta}_n - \theta_n)$ converges in distribution to a random vector $\xi$.*
2. *For any $\tau > 0$, the covariance function*

$$\begin{aligned} C_n(h_1,h_2,\Delta_\theta) &= Pf_{n,h_1,\theta_n+n^{-\nu}\Delta_\theta} f_{n,h_2,\theta_n+n^{-\nu}\Delta_\theta} \\ &\quad - Pf_{n,h_1,\theta_n+n^{-\nu}\Delta_\theta} Pf_{n,h_2,\theta_n+n^{-\nu}\Delta_\theta} \end{aligned}$$

   *converges pointwise to $C(h_1,h_2,\Delta_\theta)$ on $\mathcal{H}\times\mathcal{H}$, uniformly in $\Delta_\theta$, $\theta\in\Theta_n^\tau$.*
3. *For any $\tau > 0$, the functions $\zeta_n(h,\theta_n+n^{-\nu}\Delta_\theta)$ converges pointwise to a function $\zeta(h,\Delta_\theta)$ on $\mathcal{H}$, uniformly in $\Delta_\theta$, $\theta\in\Theta_n^\tau$.*
4. *The limiting functions $C(h_1,h_2,\Delta_\theta)$ and $\zeta(h,\Delta_\theta)$ are continuous in $\Delta_\theta$.*

*Let $Z(h,\xi)$ be a stochastic process constructed in the following manner. For a particular realization $\xi_0$ of $\xi$, generate a Gaussian process $Z(h,\xi_0)$ (independent of $\xi$) with drift $\zeta(\cdot,\xi_0)$ and covariance kernel $C(\cdot,\cdot,\xi_0)$. Then, the process $Z_n(\cdot,\hat{\theta}_n)$ converges weakly $Z(\cdot,\xi)$ in $\ell^\infty(\mathcal{H})$.*

For notational ease, we assumed each element of the vector $\hat{\theta}_n$ converges at the same rate ($n^\nu$). The extension to the general situation where different elements of $\hat{\theta}_n$ have different rates of convergence is immediate. Also, note that condition 4 is redundant when the functions $C$ and $\zeta$ do not depend on $\Delta_\theta$.

*Proof.* In light of Theorem 6.3, we only need to establish the finite dimensional convergence. Given the independence of vectors $V_i$s with $\hat{\theta}_n$, the drift process $\zeta_n(\cdot, \hat{\theta}_n)$ is independent of the centered process $(Z_n - \zeta_n)(\cdot, \hat{\theta}_n)$ given $\hat{\theta}_n$. Hence, it suffices to show the finite dimensional convergence of these two processes separately. On the set $\hat{\theta} \in \Theta_n^\tau$,

$$\begin{aligned} |\zeta_n(h, \theta_n + n^{-\nu}\Delta_{\hat{\theta}_n}) - \zeta(h, \xi)| \;\; &\leq \;\; \sup_{\theta \in \Theta_n^\tau} |\zeta_n(h, \theta_n + n^{-\nu}\Delta_\theta) - \zeta(h, \Delta_\theta)| \\ &\quad + |\zeta(h, \Delta_{\hat{\theta}_n}) - \zeta(h, \xi)|. \end{aligned}$$

In light of conditions 3 and 4, an application of Skorokhod representation theorem then ensures the convergence of finite dimensional marginals of $\zeta_n(\cdot, \theta_n + n^{-\nu}\Delta_{\hat{\theta}_n})$ to that of the process $\zeta(\cdot, \xi)$. To establish the finite dimensional convergence of the centered process $Z_n - \zeta_n$, we require the following result that arises from a careful examination of the proof of the Central Limit Theorem for sums of independent zero mean random variables (Billingsley, 1995, pp. 359 - 361).

**Theorem E.2.** *For $n \geq 1$, let $\{X_{i,n}\}_{i=1}^n$ be independent and identically distributed random variables with mean zero and variance $\sigma_n^2 > 0$. Let $S_n = (1/\sqrt{n}) \sum_{i \leq n} X_{i,n}$, $F_n$ be the distribution function of $S_n$ and for $\kappa > 0$,*

$$L_n(\kappa) = E\left[X_{1,n}^2 1\left[|X_{1,n}| > \kappa\sqrt{n}\right]\right]$$

*Then, for any $t \in \mathbb{R}$ with $|\sigma_n t| \leq \sqrt{2n}$, we have*

$$|\hat{F}_n(t) - \hat{\Phi}(\sigma_n t)| \leq \kappa\sigma_n^2|t|^3 + t^2 L_n(\kappa) + \frac{\sigma_n^4 t^4 \exp(\sigma_n^2 t^2)}{n} \tag{E.8}$$

*Here* $\hat{}$ *denotes characteristic function, so that* $\hat{\Phi}(t) = \int_{\mathbb{R}} e^{\imath tx}\Phi\{dx\}$.

We now prove Lemma E.1. Let $k \geq 1$, $c = (c_1, \dots c_k) \in \mathbb{R}^k$, $h = (h_1, \dots, h_k) \in \mathbb{R}^k$ and for $\Delta_\theta = n^\nu(\theta - \theta_n)$,

$$T_n(\Delta_\theta) = T_n(h, c, \Delta_\theta) = \sum_{j \leq k} c_j \mathbb{G}_n f_{n,h_j,\theta_n + n^{-\nu}\Delta_\theta}.$$

Note that

$$\pi_n^2(\Delta_\theta) = \mathrm{Var}(T_n(\Delta_\theta)) = \mathrm{Var}\left(\sum_{j \leq k} c_j f_{n,h_j,\theta_n + n^{-\nu}\Delta_\theta}\right).$$

converges uniformly in $\Delta_\theta$, $\theta \in \Theta_n^\tau$, to

$$\pi_0^2(\Delta_\theta) := \sum_{j_1, j_2} c_{j_1} c_{j_2} C(h_{j_1}, h_{j_2}, \Delta_\theta).$$

By Lévy continuity theorem, it suffices to show that the characteristic function

$$(c_1, \dots c_k) \mapsto E \exp\left[\imath T_n(\Delta_{\hat{\theta}_n})\right]$$

converges to $E \exp\left[\imath \pi_0(\xi) Z\right]$, where $Z$ is a standard normal random variable independent of $\xi$ and $\Delta_{\hat{\theta}_n}$. Note that

$$\begin{aligned}\left|E \exp\left[\imath T_n(\Delta_{\hat{\theta}_n})\right] - E \exp\left[\imath \pi_0(\xi) Z\right]\right| &\leq \left|E \exp\left[\imath T_n(\Delta_{\hat{\theta}_n})\right] - E \exp\left[\imath \pi_n(\Delta_{\hat{\theta}_n}) Z\right]\right| \\ &\quad + \left|E \exp\left[\imath \pi_n(\Delta_{\hat{\theta}_n}) Z\right] - E \exp\left[\imath \pi_0(\xi) Z\right]\right|.\end{aligned}$$

The right side is further bounded (up to $4\epsilon$) by

$$\begin{aligned}
&\sup_{\theta\in\Theta_n^\tau} |E\exp[\imath T_n(\Delta_\theta)] - E\exp[\imath\pi_n(\Delta_\theta)Z]| \\
&+ \sup_{\theta\in\Theta_n^\tau} |E\exp[\imath\pi_n(\Delta_\theta)Z] - E\exp[\imath\pi_0(\Delta_\theta)Z]| \qquad \text{(E.9)} \\
&+ \left|E\exp\left[\imath\pi_0(\Delta_{\hat\theta_n})Z\right] - E\exp[\imath\pi_0(\xi)Z]\right|.
\end{aligned}$$

The second term in the above display is precisely $\sup_{\theta\in\Theta_n^\tau}|\exp(-\pi_n^2(\Delta_\theta)/2) - \exp(-\pi_0^2(\Delta_\theta)/2)|$ which converges to zero. The third term converges to zero by continuous mapping theorem. To control the first term, we apply Theorem E.2. Let

$$\begin{aligned}
L_n(\kappa,\Delta_\theta) \;\; = \;\; & P\left[\left[\sum_{j\le k} c_j(f_{n,h_j,\theta_n+n^{-\nu}\Delta_\theta} - Pf_{n,h_j,\theta_n+n^{-\nu}\Delta_\theta})\right]^2 \times\right. \\
& \left. 1\left[\left|\sum_{j\le k} c_j(f_{n,h_j,\theta_n+n^{-\nu}\Delta_\theta} - Pf_{n,h_j,\theta_n+n^{-\nu}\Delta_\theta})\right| > \sqrt{n}\kappa\right]\right].
\end{aligned}$$

Then, by Theorem E.2, the first term in (E.9) is bounded by

$$\sup_{\theta\in\Theta_n^\tau}\left[\kappa\pi_n^2(\Delta_\theta) + L_n(\kappa,\Delta_\theta) + \frac{\pi_n^4(\Delta_\theta)\exp(\pi_n^2(\Delta_\theta))}{n}\right]$$

whenever $\sup_{\theta\in\Theta_n^\tau}|\pi_n(\Delta_\theta)| \le 2\sqrt{n}$, which happens eventually as the right side is $O(1)$. To see this, note that

$$\left|\sum_{j\le k} c_j f_{n,h_j,\theta_n+n^{-\nu}\Delta_\theta}\right| \le 2k\max_j(|c_j|\vee 1)F_{n,\theta}. \qquad \text{(E.10)}$$

Then, by (6.7), $\sup_{\theta\in\Theta_n^\tau}|\pi_n(\Delta_\theta)| \le 2k\max_j(|c_j|\vee 1)\sup_{\theta\in\Theta_n^\tau} PF_{n,\theta}^2 = O(1)$. Further, using (E.10),

$$L_n(\kappa,\Delta_\theta) \le \left(2k\max_j(|c_j|\vee 1)\right)^2 P\left[\left[F_{n,\theta}^2 + PF_{n,\theta}^2\right] 1\left[F > \frac{\sqrt{n}\kappa}{\max_j(|c_j|\vee 1)} - PF_{n,\theta}\right]\right],$$

which converges to zero uniformly in $\theta \in \Theta_n^\tau$ due to conditions (6.7) and (6.8). Hence,

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta_n^\tau} |E\exp[\imath T_n(\Delta_\theta)] - E\exp[\imath\pi_n(\Delta_\theta)Z]| \quad \le \quad \kappa \limsup_{n\to\infty} \sup_{\theta\in\Theta_n^\tau} \pi_n^2(\Delta_\theta).$$

As $\sup_{\theta\in\Theta_n^\tau} \pi_n^2(\Delta_\theta) = O(1)$ and $\kappa > 0$ is arbitrary, we get the result. $\square$

## E.5 Proof of Proposition 6.7

We show that the result holds for $h > 0$. The case $h < 0$ can be shown analogously. In what follows, the dependence on $h$ is suppressed in the notations for convenience.

To start with, note that $\xi_n = n^\nu(\hat{d}_1 - d_0)$ is $O_p(1)$ and it converges in distribution to a tight random variable $\xi$ with a continuous bounded density on $\mathbb{R}$. In particular, $P\left[|\xi_n| < \delta, |\xi_n| > K_{\delta/2}\right]$ converges to $P\left[|\xi| < \delta, |\xi| > K_{\delta/2}\right] \le C\delta$, for some $C > 0$.

For $u \in \mathbb{R}$, let $F_{n_2}^u$ denote the distribution function of $T_{n_2}(u)$, where

$$T_{n_2}(u) = Z_{n_2}(h, \alpha_n, \beta_n, d_0 + un^{-\nu}) - Z_{n_2}(h, \alpha_n, \beta_n, d_0).$$

Also, let $\pi_{n_2}^2 := \pi_{n_2}^2(u) = \mathrm{Var}[T_{n_2}(u)]$. Conditional on $\xi_n = u$, $T_{n_2}$ is distributed as $T_{n_2}(u)$. Also, let $\hat{}$ denote characteristic function, so that $\hat{\Phi}(t) = \int_{\mathbb{R}} e^{\imath tx}\Phi\{dx\}$. By Lévy continuity theorem, it suffices to show that for any $t \in \mathbb{R}$,

$$E\left[\exp\left(\imath t T_{n_2}\right)\right] - \hat{\Phi}(t\pi_0)$$

converges to zero. Note that

$$\begin{aligned}
\left|E\left[\exp\left(\imath t T_{n_2}\right)\right] - \hat{\Phi}(t\pi_0)\right| &= \left|E\left[E\left[\exp\left(\imath t T_{n_2}\right) - \hat{\Phi}(t\pi_0)\middle| \xi_n\right]\right]\right| \\
&= \sup_{\delta\le|u|\le K_{\delta/2}} \left|\hat{F}_{n_2}^u(t) - \hat{\Phi}(t\pi_0)\right| + 2P\left[|\xi| < \delta, |\xi| > K_{\delta/2}\right] \\
&= \sup_{\delta\le|u|\le K_{\delta/2}} \left|\hat{F}_{n_2}^u(t) - \hat{\Phi}(t\pi_{n_2}(u))\right|
\end{aligned}$$

$$+ \sup_{\delta \le |u| \le K_{\delta/2}} \left| \hat{\Phi}(t\pi_{n_2}(u)) - \hat{\Phi}(t\pi_0) \right| + C\delta. \tag{E.11}$$

We first show that $\pi_{n_2}(u)$ converges to $\pi_0$ uniformly over $u$, $\delta \le |u| \le K_{\delta/2}$ which will ensure that the second term on the right side of the above display converges to zero. To show this, note that

$$\begin{aligned}
& T_{n_2}(u) \\
= & \ \frac{1}{n_2^{\xi}} \sum_{i=1}^{n_2} \left( \frac{\beta_n - \alpha_n}{2} + \epsilon_i \right) \left[ 1 \left[ U_i K {n_1}^{-\gamma} \in (-un^{-\nu}, -un^{-\nu} + hn^{-\eta}] \right] \right. \\
& \left. -1 \left[ U_i K {n_1}^{-\gamma} \in (0, hn^{-\eta}] \right] \right] \\
= & \ \frac{1}{n_2^{\xi}} \sum_{i=1}^{n_2} \left( \frac{\beta_n - \alpha_n}{2} + \epsilon_i \right) \left[ 1 \left[ U_i K p^{-\gamma} \in (-un^{-\nu+\gamma}, -un^{-\nu+\gamma} + hn^{-\nu}] \right] \right. \\
& \left. -1 \left[ U_i K p^{-\gamma} \in (0, hn^{-\nu}] \right] \right].
\end{aligned}$$

Hence, $\pi_{n_2}$ can be simplified as

$$\begin{aligned}
& \pi_{n_2}^2(u) = \mathrm{Var}[T_{n_2}(u)] \\
= & \ \frac{n_2}{n_2^{2\xi}} E \left[ ((\beta_n - \alpha_n)/2 - \epsilon) \left[ 1 \left[ UKp^{-\gamma} \in (-un^{-\nu+\gamma}, -un^{-\nu+\gamma} + hn^{-\nu}] \right] \right. \right. \\
& \left. \left. -1 \left[ UKp^{-\gamma} \in (0, hn^{-\nu}] \right] \right] \right]^2 \\
= & \ \frac{n_2}{n_2^{2\xi}} E \left[ \left( (\beta_n - \alpha_n)^2/4 + \sigma^2 \right) \times \right. \\
& \left. 1 \left[ UKp^{-\gamma} \in (-un^{-\nu+\gamma}, -un^{-\nu+\gamma} + hn^{-\nu}] \triangle (0, hn^{-\nu}] \right] \right].
\end{aligned}$$

For $n > N_1 = (h/|\delta|)^{1/\nu}$, the sets $(-un^{-\nu+\gamma}, -un^{-\nu+\gamma} + hn^{-\nu}]$ and $(0, hn^{-\nu}]$ are disjoint and hence,

$$\pi_{n_2}^2(u) \ = \ \frac{n_2}{n_2^{2\xi}} \left( \frac{c_0^2}{4} n^{-2\xi} + \sigma^2 \right) \left[ \frac{2hn^{-\nu}}{2Kp^{-\gamma}} \right] = \pi_0^2 + \tilde{C} n^{-2\xi}, \tag{E.12}$$

where $\tilde{C} = c_0^2 (1-p)^{1-2\xi} h/(4K)$. Consequently, $\pi_{n_2}^2(u)$ converges to $\pi_0^2$ uniformly over $u$.

Next, we apply Theorem E.2 to show that the first term in (E.11) converges to zero. Write $T_{n_2}(h)$ as $(1/\sqrt{n_2})\sum_{i\leq n_2} R_{i,n_2}(u)$, where

$$\begin{aligned} R_{i,n_2}(u) &= n_2^{1/2-\xi}\left(\frac{\beta_n-\alpha_n}{2}+\epsilon_i\right)\left[1\left[U_iKp^{-\gamma}\in(-un^{-\nu+\gamma},-un^{-\nu+\gamma}+hn^{-\nu}]\right]\right. \\ &\quad \left.-1\left[U_iKp^{-\gamma}\in(0,hn^{-\nu}]\right]\right]. \end{aligned}$$

As $\gamma<\nu$, the intervals $(-un^{-\nu+\gamma},-un^{-\nu+\gamma}+hn^{-\nu}]$ and $(0,hn^{-\nu}]$ are both contained in $[-Kp^{-\gamma},Kp^{-\gamma}]$ for $n>N_2=\max\left\{(K_{\delta/2}/Kp^{-\gamma})^{1/(\nu-\gamma)},(h/Kp^{-\gamma})^{1/\nu}\right\}$ and have the same Lebesgue measure $hn^{-\nu}$. Hence, $E[T_{n_2}(u)]=E[R_{i,n_2}(u)]=0$ for $n>N_1$. Thus $T_{n_2}(u)$ is a normalized sum of mean zero random variables. Let

$$L_{n_2}(\kappa,u)=E\left[R_{i,n_2}(u)^2 1\left[|R_{i,n_2}(u)|>\sqrt{n_2}\kappa\right]\right]. \tag{E.13}$$

Using Theorem E.2, for any $\kappa>0$, $n_2>\max(N_1,N_2)$ and $|\pi_{n_2}(u)t|\leq\sqrt{2n_2}$ (which holds eventually) we have

$$|\hat{F}^u_{n_2}(t)-\hat{\Phi}(\pi_{n_2}(u)t)|\leq\kappa\pi^2_{n_2}(u)|t|^3+t^2L_{n_2}(\kappa,u)+\frac{\pi^4_{n_2}(u)t^4\exp(\pi^2_{n_2}(u)t^2)}{n_2} \tag{E.14}$$

As $\sup_{\delta\leq|u|\leq K_{\delta/2}}\pi_{n_2}(u)=O(1)$ and $\kappa$ is arbitrary, it suffices to show that

$$\sup_{\delta\leq|u|\leq K_{\delta/2}} L_{n_2}(\kappa,u)$$

converges to zero. Using the expression for $\pi_{n_2}$ in (E.12), we have

$$\begin{aligned} &L_{n_2}(\kappa,u) \\ &\leq \frac{n_2}{n_2^{2\xi}}E\left[\epsilon^2\left[1\left[UKp^{-\gamma}\in(-un^{-\nu+\gamma},-un^{-\nu+\gamma}+hn^{-\nu}]\triangle(0,hn^{-\nu}]\right]\right]\times\right. \\ &\quad\left. 1\left[n_2^{1/2-\xi}|\epsilon|>\sqrt{n_2}\kappa\right]\right]+\tilde{C}n^{-2\xi} \\ &\lesssim n^{-2\xi}+E\,\epsilon^2 1\left[|\epsilon|>\kappa n_2^{\xi}\right], \end{aligned}$$

which converges to zero uniformly in $u$. Hence, the first term in right side of (E.11) converges to zero. As $\delta > 0$ is arbitrary, we get the result. $\square$

## E.6 Proof of Theorem 6.10

We derive bounds in terms of $n$ ($n_1$, $n_2$ and $n$ have the same order). Firstly, note that $0 \in \mathcal{D}_\theta$, for all $\theta \in \Theta^\tau_{n_1}$, whenever $n > N^{(1)}_\tau := (1/p)(K_\tau/K)^{3/(1-3\gamma)}$. Further, as $r'(d_0) > 0$ and $r'$ is continuously differentiable, there exists $\delta_0 > 0$ such that $|r'(x) - r'(d_0)| < r'(d_0)/2$ (equivalently, $r'(d_0)/2 < r'(x) < 3r'(d_0)/2$) for $x \in [d_0 - \delta_0, d_0 + \delta_0]$. As $u \in \mathcal{D}_\theta$ and $\theta \in \Theta^\tau_{n_1}$, $|d_0 + un_2^{-\gamma}| < K_\tau n_1^{-1/3} + Kn_1^{-\gamma} < \delta_0$ for $n > N^{(2)}_{\tau,\delta_0} := (1/p)((K_\tau + K)/\delta_0)^{1/\gamma}$. Hence, for $n > N^{(3)}_{\tau,\delta_0} := \max(N^{(1)}_\tau, N^{(2)}_{\tau,\delta_0})$, by a change of variable,

$$\begin{aligned} M_{n_2}(u,\theta) &= n_2^\gamma \left[ \int_{d_0}^{d_0+un_2^{-\gamma}} (r(t) - r(d_0)) \frac{n_1^\gamma}{2K} dt \right] \\ &\geq n_2^\gamma \left[ \int_{d_0}^{d_0+un_2^{-\gamma}} \frac{r'(d_0)}{2} (t - d_0) \frac{n_1^\gamma}{2K} dt \right] \gtrsim u^2 =: \rho^2_{n_2}(u,0). \end{aligned}$$

Using Theorem 6.1, we need to bound

$$\sup_{\theta \in \Theta^\tau_n} E^* \sup_{|u| \leq \delta, u \in \mathcal{D}_\theta} |(\mathbb{M}_{n_2}(u,\theta) - M_{n_2}(u,\theta)) - (\mathbb{M}_{n_2}(0,\theta) - M_{n_2}(0,\theta))| \tag{E.15}$$

Recall that $\mathbb{M}_{n_2}(0,\theta) = M_{n_2}(0,\theta) = 0$. Also,

$$\sqrt{n}|\mathbb{M}_{n_2}(u,\theta) - M_{n_2}(u,\theta)| = |\mathbb{G}_{n_2} g_{n_2,u,\theta}|$$

The class of functions $\mathcal{F}_{\delta,\theta} = \{g_{n_2,u,\theta} : |u| \leq \delta, u \in \mathcal{D}_\theta\}$ is a VC class of index at most 3, with a measurable envelope (for $n > N^{(3)}_{\tau,\delta_0}$)

$$\begin{aligned} M_{\delta,\theta} &= n_2^\gamma (2\|r\|_\infty + |\epsilon|) \times \\ &\quad 1\left[UKn_1^{-\gamma} \in \left[d_0 - \theta - \delta n_2^{-\gamma}, d_0 - \theta + \delta n_2^{-\gamma}\right]\right]. \end{aligned}$$

Note that

$$E\left[M_{\delta,\theta}\right]^2 \lesssim n_2^\gamma P\left[UKn_1^{-\gamma} \in \left[d_0 - \theta - \delta n_2^{-\gamma}, d_0 - \theta + \delta n_2^{-\gamma}\right]\right] \lesssim \delta.$$

Further, the uniform entropy integral for $\mathcal{F}_{\delta,\theta}$ is bounded by a constant which only depends upon the VC-indices, i.e., the quantity

$$J(1, \mathcal{F}_{\delta,\theta}) = \sup_Q \int_0^1 \sqrt{1 + \log N(u\|M_{\delta,\theta}\|_{Q,2}, \mathcal{F}_{\delta,\theta}, L_2(Q))}du$$

is bounded. Using Theorem 2.14.1 of van der Vaart and Wellner (1996), we have

$$E^* \sup_{|u| \leq \delta u \in \mathcal{D}_\theta} n_2^\gamma \left|\mathbb{G}_{n_2} g_{n_2,u,\theta}\right| \lesssim J(1, \mathcal{F}_{\delta,\theta})\|M_{\delta,\theta}\|_2 \lesssim \delta^{1/2}.$$

Note that this bound is uniform in $\theta \in \Theta_n^\tau$. Hence, a candidate for $\phi_n(\cdot)$ to apply Theorem 6.1 is $\phi_n(\delta) = \delta^{1/2}$. The sequence $r_n = n^{(1-2\gamma)/3}$ satisfies the conditions $r_n^2 \phi_n(1/r_n) \leq \sqrt{n_2}$. As a consequence, $r_n \hat{u} = O_p(1)$. $\square$

## E.7 Proof of Theorem 6.11

We outline the main steps of the proof below. Note that

$$\begin{aligned} f_{n_2,w,\theta} &= n_2^{1/6-\gamma/3}(r(\theta + UKn_1^{-\gamma}) + \epsilon - r(d_0)) \times \\ &\quad \left(1\left[\theta + UKn_1^{-\gamma} \leq d_0 + wn_2^{-(\alpha+\gamma)}\right] - 1\left[\theta + UKn_1^{-\gamma} \leq d_0\right]\right). \end{aligned}$$

For any $L > 0$, we use Theorem 6.3 to justify the tightness of $Z_{n_2}(w, \hat{\theta}_{n_1})$ for $w \in [-L, L]$. For sufficiently large $n$, the set $\{w : w/n_2^{\alpha} \in \mathcal{D}_\theta\}$ contains $[-L, L]$ for all $\theta \in \Theta_{n_1}^{\tau}$ and hence, it is not necessary to extend $Z_{n_2}$ (equivalently, $f_{n_2,w,\theta}$) as done in (6.5). For a fixed $\theta \in \Theta_{n_1}^{\tau}$ and an envelope for $\{f_{n_2,w,\theta} : w \in [-L, L]\}$ is given by

$$F_{n_2,\theta}(V) = n_2^{1/6-\gamma/3}(2\|r\|_\infty + |\epsilon|)1\left[\theta + UKn_1^{-\gamma} \in [d_0 - Ln_2^{-(\alpha+\gamma)}, d_0 + Ln_2^{-(\alpha+\gamma)}]\right].$$

Further, $PF_{n,\theta}^2 \lesssim n^{1/3-2\gamma/3}n^{-\alpha} = O(1)$. Also,

$$P\left[F_{n_2,\theta}^2 1[F_{n_2,\theta} > \sqrt{n_2}t]\right] \lesssim E\epsilon^2 1\left[2\|r\|_\infty + |\epsilon| > \sqrt{n_2}n^{-1/6+\gamma/3}t\right],$$

which goes to zero (uniformly in $\theta$) as $E[\epsilon^2] < \infty$. Hence, conditions (6.7) and (6.8) of Theorem 6.3 are verified. With $\tilde{\rho}(w_1, w_2) = |w_1 - w_2|$, conditions (6.9) and (6.10) can be justified by elementary calculations. We justify (6.10) below. For $-L \le w_2 \le w_1 \le L$ and sufficiently large $n$ (such that $(K_\tau n_1^{-1/3} + Ln_2^{-(1+\gamma)/3}) < \min(Kn_1^{-\gamma}, \delta_0)$ with $\delta_0$ as defined in the proof of Theorem 6.10), a change of variable and boundedness of $r'$ in a $\delta_0$-neighborhood of $d_0$ yields

$$\begin{aligned}
|\zeta_{n_2}(w_1, \theta) - \zeta_{n_2}(w_2, \theta)| &\le n_2^{2/3-\gamma/3} \int_{d_0+w_2 n_2^{-(1+\gamma)/3}}^{d_0+w_1 n_2^{-(1+\gamma)/3}} (r(s) - r(d_0))\frac{n_1^{\gamma}}{2K} ds \\
&= n_2^{1/3-2\gamma/3} \int_{w_2}^{w_1} (r(d_0 + tn_2^{-(1+\gamma)/3}) - r(d_0))\frac{n_1^{\gamma}}{2K} ds \\
&\lesssim \frac{3r'(d_0)}{4}(w_1 - w_2)^2.
\end{aligned}$$

The above bound does not involve $\theta$ and converges to zero when $|w_1 - w_2|$ goes to zero. Hence, condition (6.10) holds.

Further, for a fixed $\theta$, the class $\{f_{n_2,w,\theta} : w \in [-L, L]\}$ is VC of index at most 3 with envelope $F_{n,\theta}$. Hence, the entropy condition in (6.11) is satisfied. The measurability condition (6.13) can be readily justified as well. Hence, the processes $Z_{n_2}$ are asymptotically tight for $w$ in any fixed compact set.

For a fixed $\theta \in \Theta_n^\tau$, $w \in [0, L]$ and sufficiently large $n$, $\zeta_{n_2}(w, \theta)$ equals

$$
\begin{aligned}
& n_2^{2/3-\gamma/3} \int_{d_0}^{d_0+wn_2^{-(1+\gamma)/3}} (r(s) - r(d_0)) \frac{n_1^\gamma}{2K} ds \\
= \; & \frac{(1-p)^{2/3-\gamma/3} p^\gamma n^{2/3+2\gamma/3}}{2K} \int_{d_0}^{d_0+wn_2^{-(1+\gamma)/3}} (r(s) - r(d_0)) ds \\
= \; & \frac{(1-p)^{2/3-\gamma/3} p^\gamma n^{1/3+\gamma/3}}{2K(1-p)^{(1+\gamma)/3}} \int_0^w (r(d_0 + tn_2^{-(1+\gamma)/3}) - r(d_0)) dt \\
= \; & \frac{(1-p)^{-\gamma} p^\gamma}{2K} \frac{r'(d_0)}{2} w^2 + o(1).
\end{aligned}
$$

This convergence is uniform in $\theta$ by arguments paralleling those for justifying condition (6.10).

Note that $Pf_{n_2,w,\theta} = \zeta_{n_2}(w, \theta)/\sqrt{n_2}$ converges to zero. Hence, for a fixed $\theta \in \Theta_n^\tau$ and $w_1, w_2 \in [0, L], L > 0$, the covariance function of $Z_{n_2}$ eventually equals (up to an $o(1)$ term which does not depend on $\theta$ due to a change of variable)

$$
\begin{aligned}
& P\left[f_{n_2,w_1,\theta} f_{n_2,w_2,\theta}\right] \\
= \; & n_2^{1/3-2\gamma/3} \int_0^{(w_1 \wedge w_2)n_2^{-(1+\gamma)/3}} \left[\sigma^2 + (r(d_0 + s) - r(d_0))^2\right] \frac{n_1^\gamma}{2K} ds \\
= \; & \frac{p^\gamma n^{1/3+\gamma/3}}{2K(1-p)^{-1/3+2\gamma/3}} \times \\
& \int_0^{(w_1 \wedge w_2)n_2^{-(1+\gamma)/3}} \left[\sigma^2 + (r(d_0 + s) - r(d_0))^2\right] ds \\
= \; & \frac{p^\gamma}{2K(1-p)^\gamma} \int_0^{(w_1 \wedge w_2)} \left[\sigma^2 + (r(d_0 + tn_2^{-(1+\gamma)/3}) - r(d_0))^2\right] ds
\end{aligned}
$$

$$= \frac{p^\gamma}{2K(1-p)^\gamma}(w_1 \wedge w_2)\sigma^2 + o(1).$$

This justifies the form of the limit process $Z$. Note that the process $Z \in C_{min}(\mathbb{R})$ with probability one (using argmin versions of Lemmas 2.5 and 2.6 of Kim and Pollard (1990)) and it possesses a unique argmin almost surely which is tight (the Chernoff random variable). An application of argmin continuous mapping theorem (Kim and Pollard, 1990, Theorem 2.7) along with (6.25) and rescaling arguments gives us the result. $\square$

## E.8 Proof of Theorem 6.14

Let $M(d) = P\left[Y^{(1)} 1\left[|X^{(1)} - d| < b\right]\right]$. For $F(t) = \int_0^t m(x + d_0)dx$, we have

$$M(d) = F(d - d_0 + b) - F(d - d_0 - b).$$

Note that $M'(d) = 0$ implies $m(d + b) = m(d - b)$ which holds for $d = d_0$. Hence, $d_0$ maximizes $M(\cdot)$. Also, note that $M''(d_0) = m'(d_0+b) - m'(d_0-b) = 2m'(d_0+b) < 0$. For $d$ in a small neighborhood of $d_0$ (such that $d+b > d_0$ and $2m'(d+b) \leq m'(d_0+b)$), we get

$$M(d) - M(d_0) \leq -|m'(d_0 + b)|(d - d_0)^2.$$

Note that we derived an upper bound here as our estimator is an argmax (instead of an argmin) of the criterion $\mathbb{M}_{n_1}$. Hence, the distance for applying Theorem 3.2.5 of van der Vaart and Wellner (1996) can be taken to be $\rho(d, d_0) = |d - d_0|$. The consistency of $\hat{d}_1$ with respect to $\rho$ can be deduced through standard Glivenko-Cantelli arguments and an application of argmax continuous mapping theorem (van der Vaart and Wellner, 1996, Corollary 3.2.3). For sufficiently small $\delta > 0$, consider the modulus of continuity

$$\begin{aligned} & E^* \sup_{|d-d_0|<\delta} \sqrt{n_1}|(\mathbb{M}_{n_1} - M)(d) - (\mathbb{M}_{n_1} - M)(d_0)| \\ = \;& E^* \sup_{|d-d_0|<\delta} \left|\mathbb{G}_{n_1} Y^{(1)} \left\{1\left[|X^{(1)} - d| \le b\right] - 1\left[|X^{(1)} - d_0| \le b\right]\right\}\right| \end{aligned}$$

An envelope for the class of functions $\mathcal{F}_\delta = \{g_d(x,y) = y\{1[|x-d| \le b] - 1[|x-d_0| \le b]\} : |d-d_0| < \delta\}$ is given by

$$F_\delta(X^{(1)}, \epsilon) = (\|m\|_\infty + |\epsilon|) 1\left[|X^{(1)} - d_0| \in [b-\delta, b+\delta]\right].$$

Note that $\|F_\delta\|_2 \lesssim \delta^{1/2}$. Further, the uniform entropy integral for $\mathcal{F}_\delta$ is bounded by a constant which only depends upon the VC-indices, i.e., the quantity

$$J(1, \mathcal{F}_\delta) = \sup_Q \int_0^1 \sqrt{1 + \log N(u\|F_\delta\|_{Q,2}, \mathcal{F}_\delta, L_2(Q))} du$$

is bounded. Using Theorem 2.14.1 of van der Vaart and Wellner (1996), we have

$$E^* \sup_{|d-d_0|<\delta} \sqrt{n_1}|(\mathbb{M}_{n_1} - M)(d) - (\mathbb{M}_{n_1} - M)(d_0)| \lesssim J(1, \mathcal{F}_\delta)\|F_\delta\|_2 \lesssim \delta^{1/2}.$$

Hence, a candidate for $\phi_n(\delta)$ in Theorem 3.2.5 of van der Vaart and Wellner (1996) is $\phi_n(\delta) = \delta^{1/2}$. This yields $n_1^{1/3}(\hat{d}_1 - d_0) = O_p(1)$. Next, consider the local process,

$$Z_{n_1}(h) = n_1^{2/3} \mathbb{P}_{n_1} Y^{(1)} \left[1\left[|X^{(1)} - (d_0 + hn_1^{-1/3})| < b\right] - 1\left[|X^{(1)} - d_0| < b\right]\right].$$

Note that

$$\begin{aligned} E[Z_{n_1}(h)] &= n_1^{2/3}\left\{M(d_0 + hn_1^{-1/3}) - M(d_0)\right\} \\ &= \frac{M''(d_0) + o(1)}{2}(hn_1^{-1/3})^2 n_1^{2/3} \end{aligned}$$

$$= \quad m'(d_0+b)h + o(1) = -ch + o(1).$$

Let $G(t) = \int_0^t m^2(d_0+x)dx$. Then,

$$\begin{aligned}
&\mathrm{Var}(Z_{n_1}(h)) \\
&= \frac{n_1^{4/3}}{n_1^2}\,\mathrm{Var}\left[Y^{(1)}\left[1\left[|X^{(1)}-(d_0+hn_1^{-1/3})|<b\right]-1\left[|X^{(1)}-d_0|<b\right]\right]\right] \\
&= n_1^{1/3}E\left[(Y^{(1)})^2\left[1\left[|X^{(1)}-(d_0+hn_1^{-1/3})|<b\right]-1\left[|X^{(1)}-d_0|<b\right]\right]^2\right] \\
&\quad +o(1) \\
&= n_1^{1/3}\left[G(b+hn_1^{-1/3})-G(b)+G(-b+hn_1^{-1/3})-G(-b)+2\sigma^2hn_1^{-1/3}\right] \\
&= (m^2(d_0+b)+m^2(d_0-b)+2\sigma^2)h+o(1) \\
&= 2(m^2(d_0+b)+\sigma^2)h+o(1) = a^2h+o(1).
\end{aligned}$$

The limiting covariance function can be derived in an analogous manner and the tightness of the process follows from an application of Theorem 2.11.22 of van der Vaart and Wellner (1996) involving routine justifications. An application of argmax continuous mapping theorem (van der Vaart and Wellner, 1996, Theorem 3.2.2) gives

$$n_1^{1/3}(\hat{d}_1-d_0) \xrightarrow{d} \mathrm{argmax}\left\{aB(h)-ch^2\right\}.$$

By rescaling arguments, we get the result. □

## E.9 Proof of Theorem 6.16

*Rate of convergence.* As $\gamma < 1/3$, for all $\theta \in \Theta_{n_1}^{\tau} = [\theta_0 - K_\tau n_1^{-1/3}, \theta_0 + K_\tau n_1^{-1/3}]$, $d_0 \in \mathcal{D}_\theta$, whenever $n > N_\tau^{(1)} := (1/p)(K_\tau/(K-b))^{3/(1-3\gamma)}$. For $d \in \mathcal{D}_\theta$, the set $\{u : |\theta + uKn_1^{-\gamma} - d| \le bn_1^{-\gamma}\} \subset [-1,1]$. Hence, by a change of variable,

$$M_{n_2}(d,\theta) \quad := \quad E\left[\mathbb{M}_{n_2}(d,\theta)\right]$$

$$
\begin{aligned}
&= \frac{1}{2}\int_{-1}^{1} m(\theta + uKn_1^{-\gamma}) 1\left[|\theta + uKn_1^{-\gamma} - d| \le bn_1^{-\gamma}\right] du \\
&= \frac{1}{2}\int_{\mathbb{R}} m(\theta + uKn_1^{-\gamma}) 1\left[|\theta + uKn_1^{-\gamma} - d| \le bn_1^{-\gamma}\right] du \\
&= \frac{n_1^{\gamma}}{2K}\int_{\mathbb{R}} m(x) 1\left[|x - d| \le bn_1^{-\gamma}\right] dx \\
&= \frac{n_1^{\gamma}}{2K}\int_{d-bn_1^{-\gamma}}^{d+bn_1^{-\gamma}} m(x)dx.
\end{aligned}
$$

Let

$$
F_n(d) = \int_{d-bn_1^{-\gamma}}^{d+bn_1^{-\gamma}} m(x)dx.
$$

Note that $F_n'(d) = m(d + bn_1^{-\gamma}) - m(d - bn_1^{-\gamma})$. Also,

$$
\begin{aligned}
F_n''(d) &= m'(d + bn_1^{-\gamma}) - m'(d - bn_1^{-\gamma}) \\
&= m'(d + bn_1^{-\gamma}) + m'(2d_0 - d + bn_1^{-\gamma}),
\end{aligned}
$$

whenever $d \neq d_0 \pm bn_1^{-\gamma}$. Here, the last step follows from the anti-symmetry of $m'$ around $d_0$ (but not at $d_0$). Further, as $-m'(d_0+) > 0$ and $\tilde{m}$ is continuously differentiable in a neighborhood of 0, there exists $\delta_0 > 0$ such that $|m'(x) - m'(d_0+)| < -m'(d_0+)/2$ (equivalently, $3m'(d_0+)/2 < m'(x) < m'(d_0+)/2$) for $x \in (d_0, d_0 + \delta_0]$. For $d \in \mathcal{D}_\theta$ and $\theta \in \Theta_{n_1}^{\tau}$, $|d \pm bn_1^{-\gamma} - d_0| < K_\tau n_1^{-1/3} + Kn_1^{-\gamma} < \delta_0$ for $n > N_{\tau,\delta_0}^{(2)} := (1/p)((K_\tau + K)/\delta_0)^{1/\gamma}$. Let $\rho_n^2(d, d_0) = n_1^{\gamma}(d - d_0)^2$. For $n > N_{\tau,\delta_0}^{(3)} := \max(N_\tau^{(1)}, N_{\tau,\delta_0}^{(2)})$ and $\rho_n(d, d_0) < \kappa_n := bn_1^{-\gamma/2}$ (so that $d_0 \in [d - bn_1^{-\gamma}, d + bn_1^{-\gamma}]$),

$$
\begin{aligned}
F_n''(d) &= m'(d + bn_1^{-\gamma}) + m'(2d_0 - d + bn_1^{-\gamma}) \\
&\le 2(-m'(d_0+)/2) = m'(d_0+) = -|m'(d_0+)|.
\end{aligned}
$$

Consequently, by a second order Taylor expansion,

$$\begin{aligned} M_{n_2}(d,\theta) - M_{n_2}(d_0,\theta) &= \frac{n_1^{\gamma}}{2K}\left[F_n(d) - F_n(d_0)\right] \\ &\leq -\frac{n_1^{\gamma}}{2K}\frac{|m'(d_0+)|}{2}(d-d_0)^2 \\ &\lesssim -n_1^{\gamma}(d-d_0)^2 = (-1)\rho_n^2(d,d_0). \end{aligned}$$

Again, an upper bound is deduced here as we are working with an argmax estimator.

**Claim A.** We claim that $P\left[\rho_n(\hat{d}_n, d_0) \geq \kappa_n\right]$ converges to zero. We first use the claim to prove the rate of convergence. To apply Theorem 6.1, we need to bound

$$\sup_{\theta\in\Theta_{n_1}^{\tau}} E^* \sup_{\substack{|d-d_0|<n_1^{-\gamma/2}\delta \\ d\in\mathcal{D}_\theta}} \sqrt{n_2}\left|(\mathbb{M}_{n_2}(d,\theta) - M_{n_2}(d,\theta)) - (\mathbb{M}_{n_2}(d_0,\theta) - M_n(d_0,\theta))\right|. \tag{E.16}$$

Note that

$$\sqrt{n_2}\left((\mathbb{M}_{n_2}(d,\theta) - M_{n_2}(d,\theta)) - (\mathbb{M}_{n_2}(d_0,\theta) - M_n(d_0,\theta))\right) = \mathbb{G}_{n_2} g_{n_2,d,\theta}(V),$$

where

$$\begin{aligned} g_{n_2,d,\theta}(V) &= \left[m(\theta + UKn_1^{-\gamma}) + \epsilon\right] \times \\ &\quad \left[1\left[|\theta + UKn_1^{-\gamma} - d| < bn_1^{-\gamma}\right] - 1\left[|\theta + UKn_1^{-\gamma} - d_0| < bn_1^{-\gamma}\right]\right]. \end{aligned}$$

The class of functions $\mathcal{F}_{\delta,\theta} = \{g_{n_2,d,\theta} : |d-d_0| < n_1^{-\gamma/2}\delta, d\in\mathcal{D}_\theta\}$ is VC with index at most 3 and has a measurable envelope

$$\begin{aligned} & M_{\delta,\theta}(V) \\ &= (\|m\|_\infty + |\epsilon|) \times \\ &\quad \left[1\left[bn_1^{-\gamma} - (d_0 + n_1^{-\gamma/2}\delta) < \theta_0 + UKn_1^{-\gamma} < bn_1^{-\gamma} - (d_0 - n_1^{-\gamma/2}\delta)\right]\right. \end{aligned}$$

$$+1\left[-bn_1^{-\gamma}-(d_0+n_1^{-\gamma/2}\delta)<\theta_0+UKn_1^{-\gamma}<-bn_1^{-\gamma}-(d_0-n_1^{-\gamma/2}\delta)\right]\bigg].$$

Note that $E\left[M_{\delta,\theta}(V)\right]^2 \lesssim n^{-\gamma/2}\delta$. Hence, the uniform entropy integral for $\mathcal{F}_{\delta,\theta}$ is bounded by a constant which only depends upon the VC-indices, i.e., the quantity

$$J(1,\mathcal{F}_{\delta,\theta}) = \sup_Q \int_0^1 \sqrt{1+\log N(u\|M_{\delta,\theta}\|_{Q,2},\mathcal{F}_{\delta,\theta},L_2(Q))}du$$

is bounded. Using Theorem 2.14.1 of van der Vaart and Wellner (1996), we have

$$E^* \sup_{\substack{|d-d_0|<n_1^{-\gamma/2}\delta \\ d\in\mathcal{D}_\theta}} |\mathbb{G}_{n_2} g_{n_2,d,\theta}| \leq J(1,\mathcal{F}_{\delta,\theta})\|M_{\delta,\theta}\|_2 \lesssim n^{\gamma/4}\delta^{1/2}.$$

The above bound is uniform in $\theta \in \Theta_{n_1}^{\tau}$. Hence, a candidate for $\phi_n$ to apply Theorem 6.1 is $\phi_{n_2}(\delta) = n^{\gamma/4}\delta^{1/2}$. This yields $n^{(1+\gamma)/3}(\hat{d}_2 - d_0) = O_p(1)$.

*Proof of* **Claim A**. Note that $\rho_n(d,d_0) \geq \kappa_n \Leftrightarrow |d-d_0| \geq bn_1^{-\gamma}$. Also, for such $d \in \mathcal{D}_\theta$, the bin $(d-bn_1^{-\gamma}, d+bn_1^{-\gamma})$ does not contain $d_0$ and is either completely to the right of $d_0$ or to the left (regions where $m$ is continuously differentiable). In particular, for such $d$'s with $d > d_0$ and $n > N_{\tau,\delta_0}^{(3)}$,

$$F_n'(d) \;=\; m(d+bn_1^{-\gamma}) - m(d-bn_1^{-\gamma}) \leq -(|m'(d_0+)|/2)(2bn_1^{-\gamma}) = -|m'(d_0+)|bn_1^{-\gamma}.$$

As a consequence,

$$M_{n_2}(d,\theta) - M_{n_2}(d_0+bn_1^{-\gamma},\theta) \leq (n_1^{\gamma}/2K)(-(|m'(d_0+)|bn_1^{-\gamma})|d-(d_0+bn_1^{-\gamma})|) \leq 0, \tag{E.17}$$

for $d > d_0 + bn_1^{-\gamma}$. Also, for $n > N^{(3)}_{\tau,\delta_0}$,

$$
\begin{aligned}
M_{n_2}(d_0 + bn_1^{-\gamma}, \theta) - M_{n_2}(d_0, \theta) &= \frac{n_1^{\gamma}}{2K} \left[ \int_{d_0}^{d_0+2bn_1^{-\gamma}} m(x)dx - 2 \int_{d_0}^{d_0+bn_1^{-\gamma}} m(x)dx \right] \\
&= \frac{n_1^{\gamma}}{2K} \left[ \int_{d_0+bn_1^{-\gamma}}^{d_0+2bn_1^{-\gamma}} m(x)dx - \int_{d_0}^{d_0+bn_1^{-\gamma}} m(x)dx \right] \\
&= \frac{n_1^{\gamma}}{2K} \int_{d_0}^{d_0+bn_1^{-\gamma}} (m(x + bn_1^{-\gamma}) - m(x))dx \\
&\leq \frac{n_1^{\gamma}}{2K} \int_{d_0}^{d_0+bn_1^{-\gamma}} (m'(d_0)/2)bn_1^{-\gamma})dx \leq \frac{-|m'(d_0)|b^2}{4K} n_1^{-\gamma}.
\end{aligned}
\tag{E.18}
$$

Using (E.17) and (E.18),

$$
\begin{aligned}
c_n^{\tau}(\kappa_n) &= \sup_{\theta \in \Theta_n^{\tau}} \sup_{\substack{\rho_n(d,d_n) \geq \kappa_n, d > d_0 \\ d \in \mathcal{D}_\theta}} \{M_{n_2}(d, \theta) - M_{n_2}(d_0, \theta)\} \\
&\leq \sup_{\theta \in \Theta_n^{\tau}} \sup_{\substack{\rho_n(d,d_n) \geq \kappa_n, d > d_0 \\ d \in \mathcal{D}_\theta}} \left\{M_{n_2}(d, \theta) - M_{n_2}(d_0 + bn_1^{-\gamma}, \theta)\right\} \\
&\quad + \sup_{\theta \in \Theta_n^{\tau}} \sup_{\substack{\rho_n(d,d_n) \geq \kappa_n, d > d_0 \\ d \in \mathcal{D}_\theta}} \left\{M_{n_2}(d_0 + bn_1^{-\gamma}, \theta) - M_{n_2}(d_0, \theta)\right\} \\
&\lesssim -n^{-\gamma}.
\end{aligned}
$$

Note that an upper bound is derived as we are working with argmax type estimators instead of argmins. The same upper bound can deduced for the situation $d < d_0$. Further, $\mathbb{M}_{n_2}(d, \theta) - M_{n_2}(d, \theta) = (\mathbb{P}_{n_2} - P)\tilde{g}_{n_2,d,\theta}$, where

$$
\tilde{g}_{n_2,d,\theta}(V) \;=\; \left[m(\theta + UKn_1^{-\gamma}) + \epsilon\right] 1 \left[|\theta + UKn_1^{-\gamma} - d| < bn_1^{-\gamma}\right].
$$

The class of functions $\mathcal{G}_{n_2,\theta} = \{\tilde{g}_{n_2,d,\theta} : d \in \mathcal{D}_\theta\}$ is VC of index at most 3 and is enveloped by the function

$$G_{n_2}(V) = (\|m\|_\infty + |\epsilon|)$$

with $\|G_{n_2}\|_{L_2(P)} = O(1)$. Further, the uniform entropy integral for $\mathcal{G}_{n_2,\theta}$ is bounded by a constant which only depends upon the VC-indices, i.e., the quantity

$$J(1, \mathcal{G}_{n_2,\theta}) = \sup_Q \int_0^1 \sqrt{1 + \log N(u\|G_{n_2}\|_{Q,2}, \mathcal{G}_{n_2,\theta}, L_2(Q))}du$$

is bounded. Using Theorem 2.14.1 of van der Vaart and Wellner (1996),

$$E^* \sup_{\mathcal{G}_{n_2,\theta}} |\mathbb{G}_{n_2}\tilde{g}_{n_2,d,\theta}| \lesssim J(1, \mathcal{G}_{n_2,\theta})\|G_{n_2}\|_2 = O(1), \tag{E.19}$$

where the $O(1)$ term does not depend on $\theta$ (as the envelope $G_{n_2}$ does not depend on $\theta$). Consequently, by Markov inequality,

$$\sup_{\theta\in\Theta^\tau_{n_1}} P\left[2\sup_{d\in\mathcal{D}_\theta} |\mathbb{M}_n(d,\theta) - M_n(d,\theta)| > -c^\tau_n(\kappa_n)\right] \;\; \leq \;\; \frac{O(1)}{\sqrt{n}n^{-\gamma}}.$$

As $\gamma < 1/3 < 1/2$, the right side converges to zero. Hence, **Claim A** holds.

*Limit distribution.* For deriving the limit distribution, let

$$Z_{n_2}(h,\theta) = \mathbb{G}_{n_2} f_{n_2,h,\theta}(V) + \zeta_{n_2}(h,\theta),$$

where $\zeta_{n_2}(h,\theta) = \sqrt{n_2}P\left[f_{n_2,h,\theta}(V)\right]$ and

$$f_{n_2,h,\theta}(V) \;\; = \;\; n_2^{1/6-\gamma/3}(g_{n_2,d_0+hn_2^{(1+\gamma)/3},\theta}(V) - g_{n_2,d_0,\theta}(V)).$$

Further, the asymptotic tightness of processes of the type

$$\sqrt{n_2}\mathbb{G}_{n_2}(m(\theta + UKn_1^{-\gamma}) + \epsilon)1\left[d_0 - bn_1^{-\gamma} < \theta + UKn_1^{-\gamma} \le d_0 + hn_2^{-(1+\gamma)/3} + bn_1^{-\gamma}\right] \tag{E.20}$$

can be established by arguments analogous to those in the proof of Theorem 6.11. As indicators with absolute values can be split as

$$1\left[|a_1 - a_2| \le a_3\right] = 1\left[a_1 - a_2 \le a_3\right] - 1\left[a_3 < a_1 - a_2|\right],$$

the process $Z_{n_2}$ can be broken into process of the form (E.20). As the sum of tight processes is tight, we get tightness for the process $Z_{n_2}$. Further,

$$\zeta_{n_2}(h,\theta) = n_2^{1/2+1/6-\gamma/3}\left[M_{n_2}(d_0 + hn_2^{-(1+\gamma)/3}, \theta) - M_{n_2}(d_0,\theta)\right].$$

Fix $L > 0$. For $h \in [-L, L]$ and $\theta \in \Theta_{n_1}^{\tau}$, both $d_0 + hn_2^{(1+\gamma)/3}$ and $d_0$ lie in the set $\mathcal{D}_\theta$ and hence,

$$\zeta_{n_2}(h,\theta) \quad = \quad n_2^{2/3-\gamma/3}\frac{n_1^{\gamma}}{2K}\left[F_n(d_0 + hn_2^{-(1+\gamma)/3}) - F_n(d_0)\right].$$

Note that

$$F_n''(d_0 + hn_2^{-(1+\gamma)/3}) = m'(d_0 + hn_2^{-(1+\gamma)/3} + bn_1^{-\gamma}) - m'(d_0 + hn_2^{-(1+\gamma)/3} - bn_1^{-\gamma}).$$

For any $h \in [-L, L]$, $d_0 \in [d_0 + hn_2^{-(1+\gamma)/3} - bn_1^{-\gamma}, d_0 + hn_2^{-(1+\gamma)/3} - bn_1^{-\gamma}]$ eventually and hence, $F_n''(d_0 + hn_2^{-(1+\gamma)/3}) = 2m'(d_0+) + o(1)$. Consequently,

$$\begin{aligned}\zeta_{n_2}(h,\theta) &= \frac{p^{\gamma}n_2^{2/3+2\gamma/3}}{2K(1-p)^{\gamma}}\frac{F_n''(d_0 + o(1))}{2}h^2 n_2^{-2(1+\gamma)/3} \\ &= -\frac{p^{\gamma}}{(1-p)^{\gamma}}\frac{|m'(d_0+)|}{2K}h^2 + o(1).\end{aligned}$$

Note that the above convergence is uniform in $\theta \in \Theta_{n_1}^{\tau}$ (due to a change of variable allowed for large $n$). Next, we justify the form of the limiting variance function for simplicity. The covariance function can be deduced along to same lines in a notationally tedious manner. As $P\left[f_{n_2,h,\theta}(V)\right] = \zeta_{n_2}(h,\theta)/\sqrt{n}$ converges to zero, for $\theta \in \Theta_{n_1}^{\tau}$ and $h \in [0, L]$, the variance of $Z_{n_2}(h)$ eventually equals (up to an $o(1)$ term)

$$P\left[f_{n_2,h_1,\theta}^2\right] = \frac{n_2^{1/3-2\gamma/3}}{2Kn_1^{-\gamma}} \int_{\mathbb{R}} \left(\sigma^2 + m^2(x)\right) \left[1\left[|x - d_0 + hn_2^{-(1+\gamma)/3}| \leq bn_1^{-\gamma}\right] - 1\left[|x - d_0| \leq bn_1^{-\gamma}\right]\right]^2 dx.$$

Note that

$$\begin{aligned}
&\left[1\left[|x - (d_0 + hn_2^{-(1+\gamma)/3})| \leq bn_1^{-\gamma}\right] - 1\left[|x - d_0| \leq bn_1^{-\gamma}\right]\right]^2 \\
&\qquad = 1\left[d_0 + bn_1^{-\gamma} < x \leq d_0 + hn_2^{-(1+\gamma)/3} + bn_1^{-\gamma}\right] \\
&\qquad\quad +1\left[d_0 - bn_1^{-\gamma} < x \leq d_0 + hn_2^{-(1+\gamma)/3} - bn_1^{-\gamma}\right].
\end{aligned}$$

Further,

$$\begin{aligned}
&\frac{n_2^{1/3-2\gamma/3}n_1^{\gamma}}{2K} \int_{\mathbb{R}} \left(\sigma^2 + m^2(x)\right) 1\left[d_0 + bn_1^{-\gamma} < x \leq d_0 + hn_2^{-(1+\gamma)/3} + bn_1^{-\gamma}\right] dx \\
&\qquad = \frac{p^{\gamma}n_2^{1/3+\gamma/3}}{2K(1-p)^{\gamma}}(\sigma^2 + m^2(d_0) + o(1))hn_2^{-(1+\gamma)/3} \\
&\qquad = \frac{p^{\gamma}}{2K(1-p)^{\gamma}}(\sigma^2 + m^2(d_0))h + o(1).
\end{aligned}$$

Hence, the process $Z_{n_2}$ converges weakly to the process

$$Z(h) = \sqrt{\frac{p^{\gamma}}{K(1-p)^{\gamma}}(m^2(d_0) + \sigma^2)}B(h) - \frac{p^{\gamma}}{(1-p)^{\gamma}}\frac{|m'(d_0+)|}{2K}h^2.$$

Note that $Z \in C_{min}(\mathbb{R})$ with probability one. By rescaling arguments, the result follows. $\square$

# BIBLIOGRAPHY

# BIBLIOGRAPHY


Adler, R. J. and Taylor, J. E. (2007). *Random fields and geometry.* Springer Monographs in Mathematics. Springer, New York.

Anevski, D. and Hössjer, O. (2006). A general asymptotic scheme for inference under order restrictions. *Ann. Statist.*, 34(4):1874–1930.

Bambos, N. and Michailidis, G. (2004). Queueing and scheduling in random environments. *Adv. in Appl. Probab.*, 36(1):293–317.

Banerjee, M. (2009). Inference in exponential family regression models under certain shape constraints using inversion based techniques. In *Advances in multivariate statistical methods*, volume 4 of *Stat. Sci. Interdiscip. Res.*, pages 249–271. World Sci. Publ., Hackensack, NJ.

Banerjee, M. and McKeague, I. W. (2007). Confidence sets for split points in decision trees. *Ann. Statist.*, 35(2):543–574.

Banerjee, M. and Wellner, J. A. (2005). Confidence intervals for current status data. *Scand. J. Statist.*, 32(3):405–424.

Belitser, E., Ghosal, S., and Zanten, J. v. (2013). Optimal two-stage procedures for estimating location and size of maximum of multivariate regression functions. *Ann. Statist.*

Bhattacharya, P. (1987). Maximum likelihood estimation of a change-point in the distribution of independent random variables: General multiparameter case. *J. Multivariate Anal.*, 23(2):183 – 208.

Bhattacharya, P. K. and Brockwell, P. J. (1976). The minimum of an additive process with applications to signal estimation and storage theory. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 37(1):51–75.

Bierens, H. (1987). Kernel estimators of regression functions. In *Advances in Econometrics – Fifth World Congress, Vol. I*, pages 99–144, Cambridge. Cambridge University Press.

Billingsley, P. (1968). *Convergence of probability measures.* John Wiley & Sons Inc., New York.

Billingsley, P. (1995). *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, third edition. A Wiley-Interscience Publication.

Bolthausen, E. (1982). On the central limit theorem for stationary mixing random fields. *Ann. Probab.*, 10(4):1047–1050.

Bronšteĭn, E. M. (1976). $\varepsilon$-entropy of convex sets and functions. *Sibirsk. Mat. Ž.*, 17(3):508–514, 715.

Cavalier, L. (1997). Nonparametric estimation of regression level sets. *Statistics*, 29(2):131–160.

Chen, Y. (1999). Nonparametric identification of the minimum effective dose. *Biometrics*, 55:1236–1240.

Chen, Y.-I. and Chang, Y.-M. (2007). Identification of the minimum effective dose for right-censored survival data. *Comput. Statist. Data Anal.*, 51(6):3213–3222.

Cheng, M.-Y. and Raimondo, M. (2008). Kernel methods for optimal change-points estimation in derivatives. *J. Comput. Graph. Statist.*, 17(1):56–75.

Cox, C. (1987). Threshold dose-response models in toxicology. *Biometrics*, 43(3):511–523.

Cramér, H. and Leadbetter, M. R. (1967). *Stationary and related stochastic processes. Sample function properties and their applications*. John Wiley & Sons Inc., New York.

Cressie, N. A. C. (1991). *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

Csörgő, S. and Mielniczuk, J. (1995a). Close short-range dependent sums and regression estimation. *Acta Sci. Math. (Szeged)*, 60(1-2):177–196.

Csörgő, S. and Mielniczuk, J. (1995b). Nonparametric regression under long-range dependent normal errors. *Ann. Statist.*, 23(3):1000–1014.

Delworth, T. and Knutson, T. (2000). Simulation of early 20th century global warming. *Science*, 287:2246–2250.

Dudley, R. M. (1984). A course on empirical processes. In *École d'été de probabilités de Saint-Flour, XII—1982*, volume 1097 of *Lecture Notes in Math.*, pages 1–142. Springer, Berlin.

Dümbgen, L. (1991). The asymptotic behavior of some nonparametric change-point estimators. *Ann. Statist.*, 19(3):1471–1495.

Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.*, 29(1):124–152.

Dümbgen, L. and Walther, G. (2008). Multiscale inference about a density. *Ann. Statist.*, 36(4):1758–1785.

Ermakov, M. S. (1976). On the asymptotic behavior of statistical estimates for samples having a density with singularities. *Theory Probab. Appl.*, 21:649–651.

Feller, W. (1971). *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons Inc., New York.

Ferger, D. (1999). On the uniqueness of maximizers of Markov-Gaussian processes. *Statist. Probab. Lett.*, 45(1):71–77.

Ferger, D. (2004). A continuous mapping theorem for the argmax-functional in the non-unique case. *Statist. Neerlandica*, 58(1):83–96.

Gasser, T. and Müller, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.*, 11(3):171–185.

Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73(3):625–633.

Geppetti, P. and Benemei, S. (2009). Pain treatment with opioids : achieving the minimal effective and the minimal interacting dose. *Clin. Drug. Investig.*, 29(Suppl 1):3–16.

Goldenshluger, A., Juditsky, A., Tsybakov, A. B., and Zeevi, A. (2008). Change-point estimation from indirect observations. I. Minimax complexity. *Ann. Inst. Henri Poincaré Probab. Stat.*, 44(5):787–818.

Goldenshluger, A. and Spokoiny, V. (2006). Recovering convex edges of an image from noisy tomographic data. *IEEE Trans. Inform. Theory*, 52(4):1322–1334.

Goldenshluger, A., Tsybakov, A., and Zeevi, A. (2006). Optimal change-point estimation from indirect observations. *Ann. Statist.*, 34(1):350–372.

Goldenshluger, A. and Zeevi, A. (2006). Recovering convex boundaries from blurred and noisy observations. *Ann. Statist.*, 34(3):1375–1394.

Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 539–555, Belmont, CA. Wadsworth.

Groeneboom, P. (1989). Brownian motion with a parabolic drift and Airy functions. *Probab. Theory Related Fields*, 81(1):79–109.

Hall, P., Kay, J. W., and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3):521–528.

Härdle, W. and Gasser, T. (1985). On robust kernel estimation of derivatives of regression functions. *Scand. J. Statist.*, 12(3):233–240.

Härdle, W., Park, B. U., and Tsybakov, A. B. (1995). Estimation of non-sharp support boundaries. *J. Multivariate Anal.*, 55(2):205–218.

Hartigan, J. A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.*, 82(397):267–270.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 57:1–17.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30.

Holst, U., Hossjer, O., Bjorklund, C., Ragnarson, P., and Edner, H. (1996). Locally weighted least squares kernel regression and statistical evaluation of lidar measurements. *Enivironmetrics*, 7:401–416.

Hotelling, H. (1941). Experimental determination of the maximum of a function. *Ann. Math. Statistics*, 12:20–45.

Hsu, J. C. and Berger, R. L. (1999). Stepwise confidence intervals without multiplicity adjustment for doseresponse and toxicity studies. *Journal of the American Statistical Association*, 94(446):468–482.

Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Teor. Verojatnost. i Primenen.*, 7:361–392.

Ibragimov, I. A. and Has′minskiĭ, R. Z. (1981). *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York. Asymptotic theory, Translated from the Russian by Samuel Kotz.

Ivanov, A. V. and Leonenko, N. N. (1989). *Statistical analysis of random fields*, volume 28 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht. With a preface by A. V. Skorokhod, Translated from the Russian by A. I. Kochubinskiĭ.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statistics*, 23:462–466.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.*, 18(1):191–219.

Ko, M.-H., Kim, H.-C., and Kim, T.-S. (2008). On functional central limit theorems for linear random fields with dependent innovations. *ANZIAM J.*, 49(4):533–541.

Korostelëv, A. P. (1987). Minimax estimation of a discontinuous signal. *Teor. Veroyatnost. i Primenen.*, 32(4):796–799.

Korostelëv, A. P. and Tsybakov, A. B. (1993a). *Minimax theory of image reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Korostelëv, A. P. and Tsybakov, A. B. (1993b). *Minimax theory of image reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer Series in Statistics. Springer, New York.

Koul, H. L., Qian, L., and Surgailis, D. (2003). Asymptotics of $M$-estimators in two-phase linear regression models. *Stochastic Process. Appl.*, 103(1):123–154.

Lan, Y., Banerjee, M., and Michailidis, G. (2007). Change-point estimation under adaptive sampling. Technical report, http://www.stat.lsa.umich.edu/∼moulib/cp-techrep.pdf.

Lan, Y., Banerjee, M., and Michailidis, G. (2009). Change-point estimation under adaptive sampling. *Ann. Statist.*, 37(4):1752–1791.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition.

Lifshits, M. A. (1982). On the absolute continuity of the distributions of functionals of stochastic processes. *Theory Probab. Appl.*, 27(3):600–607.

Lindvall, T. (1973). Weak convergence of probability measures and random functions in the function space $D(0, \infty)$. *J. Appl. Probability*, 10:109–121.

Loader, C. R. (1996). Change point estimation using nonparametric regression. *Ann. Statist.*, 24(4):1667–1678.

Ma, L., Guo, J., Wang, Y., Tian, Y., and Yang, Y. (2010). Ship detection by salient convex boundaries. In *Image and Signal Processing (CISP), 2010 3rd International Congress on*, volume 1, pages 202–205.

Mallik, A., Banerjee, M., and Sen, B. (2013a). Asymptotics for $p$-value based threshold estimation in regression settings. submitted, available at *http://www.stat.lsa.umich.edu/∼moulib/th-estn-reg-v2.pdf*.

Mallik, A., Banerjee, M., Sen, B., and Michailidis, G. (2013b). Asymptotics for $p$-value based threshold estimation in dose–response settings. submitted, available at *http://www.stat.columbia.edu/∼bodhi/research/th-estn-mnJuly13.pdf*.

Mallik, A., Sen, B., Banerjee, M., and Michailidis, G. (2011). Threshold estimation based on a $p$-value framework in dose-response and regression settings. *Biometrika*, 98(4):887–900.

Mallik, A. and Woodroofe, M. (2011). A central limit theorem for linear random fields. *Statist. Probab. Lett.*, 81(11):1623–1626.

Mammen, E. and Tsybakov, A. B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.*, 23(2):502–524.

Marinucci, D. and Poghosyan, S. (2001). Asymptotics for linear random fields. *Statist. Probab. Lett.*, 51(2):131–141.

Melillo, J. (1999). Climate change: warm, warm on the range. *Science*, 283:183–184.

Müller, H.-G. (1992). Change-points in nonparametric regression analysis. *Ann. Statist.*, 20(2):737–761.

Müller, H.-G. and Song, K.-S. (1996). A set-indexed process in a two-region image. *Stochastic Process. Appl.*, 62(1):87–101.

Müller, H.-G. and Song, K.-S. (1997). Two-stage change-point estimators in smooth regression models. *Statist. Probab. Lett.*, 34(4):323–335.

Neumann, M. H. (1997). Optimal change-point estimation in inverse problems. *Scand. J. Statist.*, 24(4):503–521.

Nolan, D. (1991). The excess-mass ellipsoid. *J. Multivariate Anal.*, 39(2):348–371.

Paulauskas, V. (2010). On Beveridge-Nelson decomposition and limit theorems for linear random fields. *J. Multivariate Anal.*, 101(3):621–639.

Pflug, G. C. (1983). The limiting log-likelihood process for discontinuous density families. *Z. Wahrsch. Verw. Gebiete*, 64(1):15–35.

Phillips, P. C. B. and Solo, V. (1992). Asymptotics for linear processes. *Ann. Statist.*, 20(2):971–1001.

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *Ann. Statist.*, 23(3):855–881.

Pons, O. (2003). Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *Ann. Statist.*, 31(2):442–463. Dedicated to the memory of Herbert E. Robbins.

Qiu, P. (2007). Jump surface estimation, edge detection, and image restoration. *J. Amer. Statist. Assoc.*, 102(478):745–756.

Raimondo, M. (1998). Minimax estimation of sharp change points. *Ann. Statist.*, 26(4):1379–1397.

Ranga Rao, R. (1962). Relations between weak and uniform convergence of measures with applications. *Ann. Math. Statist.*, 33:659–680.

Ranke, J., Molter, K., Stock, F., Bottin-Weber, U., Poczobutt, J., Hoffmann, J., Ondruschka, B., Filser, J., and Jastorff, B. (2004). Biological effects of imidazolium ionic liquids with varying chain lengths in acute vibrio fischeri and wst-1 cell viability assays. *Ecotoxicology and Environmental Safety*, 28:396–404.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester.

Robinson, P. M. (1997). Large-sample inference for nonparametric regression with dependent errors. *Ann. Statist.*, 25(5):2054–2083.

Rosén, B. (1969). A note on asymptotic normality of sums of higher-dimensionally indexed random variables. *Ark. Mat.*, 8:33–43 (1969).

Ruppert, D., Wand, M. P., Holst, U., and Hössjer, O. (1997). Local polynomial variance-function estimation. *Technometrics*, 39(3):262–273.

Schlesinger, M. E. and Ramankutty, N. (1994). An oscillation in the global climate system of period 65-70 years. *Nature*, 367:723–726.

Scott, C. (2011). Surrogate losses and regret bounds for cost-sensitive classification with example-dependent costs. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 153–160, New York, NY, USA. ACM.

Scott, C. and Davenport, M. (2007). Regression level set estimation via cost-sensitive classification. *IEEE Trans. Signal Process.*, 55(6, part 1):2752–2757.

Seijo, E. and Sen, B. (2011). Change-point in stochastic design regression and the bootstrap. *Ann. Statist.*, 39(3):1580–1607.

Sen, B., Banerjee, M., and Woodroofe, M. (2010). Inconsistency of bootstrap: the Grenander estimator. *Ann. Statist.*, 38(4):1953–1977.

Singh, A., Scott, C., and Nowak, R. (2009). Adaptive Hausdorff estimation of density level sets. *Ann. Statist.*, 37(5B):2760–2782.

Stahl, J. S. and Wang, S. (2005). Convex grouping combining boundary and region information. In *Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2*, ICCV '05, pages 946–953, Washington, DC, USA. IEEE Computer Society.

Tamhane, A. C. and Logan, B. R. (2002). Multiple test procedures for identifying the minimum effective and maximum safe doses of a drug. *J. Amer. Statist. Assoc.*, 97(457):293–301.

Tang, R., Banerjee, M., and Michailidis, G. (2011). A two-stage hybrid procedure for estimating an inverse regression function. *Ann. Statist.*, 39(2):956–989.

Tang, R., Banerjee, M., Michailidis, G., and Mankad, S. (2013). Two-stage plans for estimating a threshold value of a regression function. submitted, available at *http://dept.stat.lsa.umich.edu/∼moulib/techno-twostage-GM.pdf*.

Tsybakov, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.*, 25(3):948–969.

van de Geer, S. (2000). *Empirical processes in M-estimation*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer Series in Statistics. Springer-Verlag, New York.

van der Vaart, A. W. and Wellner, J. A. (2007). Empirical processes indexed by estimated functions. In *Asymptotics: particles, processes and inverse problems*, volume 55 of *IMS Lecture Notes Monogr. Ser.*, pages 234–252. Inst. Math. Statist., Beachwood, OH.

Wakimoto, R. M. and McElroy, J. L. (1986). Lidar observation of elevated pollution layers over los angeles. *J. Climate Appl. Meteor.*, 25:1583–1599.

Wang, S., Stahl, J., Bailey, A., and Dropps, M. (2007). Global detection of salient convex boundaries. *International Journal of Computer Vision*, 71(3):337–359.

Wang, Y. and Woodroofe, M. (2013). A new condition for the invariance principle for stationary random fields. *Statist. Sinica*, 23(4). To appear.

Willett, R. M. and Nowak, R. D. (2007). Minimax optimal level set estimation. *IEEE Trans. Image Process.*, 16(12):2965–2979.

Williams, D. A. (1971). A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, 27:103–117.

Wishart, J. (2009). Kink estimation with correlated noise. *J. Korean Statist. Soc.*, 38(2):131–143.

Wishart, J. and Kulik, R. (2010). Kink estimation in stochastic regression with dependent errors and predictors. *Electron. J. Stat.*, 4:875–913.

Wright, F. T. (1981). The asymptotic behavior of monotone regression estimates. *Ann. Statist.*, 9(2):443–448.

Wu, W. B., Woodroofe, M., and Mentz, G. (2001). Isotonic regression: another look at the changepoint problem. *Biometrika*, 88(3):793–804.

Zhao, O. and Woodroofe, M. (2012). Estimating a monontone trend. *Statist. Sinica*, 22:359–378.