Statistical inference in two non-standard regression problems

Emilio Seijo

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2012

ABSTRACT

Statistical inference in two non-standard regression problems

Emilio Seijo

This thesis analyzes two regression models in which their respective least squares estimators have nonstandard asymptotics. It is divided in an introduction and two parts. The introduction motivates the study of nonstandard problems and presents an outline of the contents of the remaining chapters. In part I, the least squares estimator of a multivariate convex regression function is studied in great detail. The main contribution here is a proof of the consistency of the aforementioned estimator in a completely nonparametric setting. Model misspecification, local rates of convergence and multidimensional regression models mixing convexity and componentwise monotonicity constraints will also be considered. Part II deals with change-point regression models and the issues that might arise when applying the bootstrap to these problems. The classical bootstrap is shown to be inconsistent on a simple change-point regression model, and an alternative (smoothed) bootstrap procedure is proposed and proved to be consistent. The superiority of the alternative method is also illustrated through a simulation study. In addition, a

version of the continuous mapping theorem specially suited for change-point estimators is proved and used to derive the results concerning the bootstrap.

Contents

List of Figures

Acknowledgments

I want to take this opportunity to express my most sincere gratitude to my advisor, professor Bodhisattva Sen. It would have been impossible for me to complete this thesis without his patience, guidance and direction. It was a privilege to work closely with someone who enjoys statistical inference as much as he does. I would also like to acknowledge all my professors at Columbia from whom I learned so much. I am particularly grateful to professors Richard Davis, Zhiliang Ying, David Madigan and José Blanchet. The advice they gave me at different stages during the completion of the Ph.D. program were very helpful and enlightening. In addition, I would like to thank my dissertation committee: Professor Richard Davis, professor Ian McKeague, professor Jos'é Blanchet, professor Zhiliang Ying, and professor Bodhisattva Sen. I truly appreciate the time they took to read through my work.

I am also grateful to all my classmates and friends at Columbia. I would like to acknowledge all the help that I received, especially during my first couple of years in the Ph.D. program, from Tyler McCormick, Subhankar Sadhukan, Johannes Ruf and Gerardo Hernández del Valle. My discussions with them were always illustrative and taught me a lot about probability, statistics and programming. I also want to recognize the work of all the staff at Columbia, and particularly Anthony Cruz and Dood Kalicharan. They were an integral part of my success at Columbia.

I am thankful to professor Moulinath Banerjee, from the University of Michigan, and Souvik Gosh. Their help was essential for the completion of this work.

Last but not least, I want to thank my family and friends for all their support. I am particularly grateful to my parents and Alejandra, for their ever present support and advice.

Chapter 1

Introduction

This dissertation comprises the statistical analysis of two regression models. The first of these is a regression problem in which the regressand is a convex function of a possibly multidimensional regressor. The other one is the so-called change-point regression problem, and it consists in estimating a jump discontinuity (change-point) in an otherwise smooth curve. Though quite different in nature, these problems share a common characteristic: both can be solved with least squares estimation procedures which exhibit nonstandard asymptotics.

A sequence of consistent estimators in a point estimation problem is said to have nonstandard asymptotics if the estimators converge to a non-Gaussian limiting distribution at a rate other than $n^{-1/2}$. A trivial example arises in the estimation of $\theta > 0$ given a random sample from a $Uniform(0, \theta)$ distribution. In this case, the maximum likelihood estimator (MLE), which is the maximum of the sample, converges at rate n^{-1} to an $Exponential\,(\theta^{-1})$ distribution. This problem does not satisfy the regularity conditions that are usually assumed for MLE's (see either Lehmann and Casella (1998) or van der Vaart (1998)). Thus, the standard asymptotic theory of the parametric MLE does not apply and the result has to be deduced via direct calculations. Despite its simplicity, this problem illustrates the fact that nonstandard problems require specially tailored solutions.

Nonstandard problems are frequently encountered outside the realm of parametric statistical inference and some of them have been carefully studied in the literature. For instance, Kim and Pollard (1990) show a family of cube-root asymptotic problems arising from a wide array of applications while Groeneboom et al. (2001) prove that the univariate least squares estimator in convex regression exhibits nonstandard asymptotics (see Section 3.1) In this context, this thesis presents an illustration of the issues that might arise in nonstandard problems and the techniques that can be used to deal with them.

The first part of the thesis deals with multidimensional convex regression. This problem involves the estimation of a function with a multidimensional argument subject to a shape-restriction (convexity). We will define the least squares estimator in multiple dimensions, provide means for its computations, describe its finite sample properties and prove its strong consistency (and that of its subdifferentials). This is one of the main contributions of this thesis as it constitutes the first attempt to solve this problem in a completely nonparametric setting.

In addition to the consistency of the least squares estimator in multidimensional convex regression, we will treat some other topics regarding convex function estimation. In Section 3.1 we describe the complete local asymptotic theory in the one-dimensional case, illustrating that convex regression is a nonstandard problem. In Section 3.2 we will generalize the methods of Chapter 2 to the case in which the regression function is known to be convex and monotone in some subset of the coordinates of its argument. We will argue that the least squares estimator is also consistent in this situation. In

Section 3.3 we will describe some results regarding the behavior of the least squares estimators under misspecified models. We will finish the first part of the thesis by providing a conjecture about local rates of convergence for the least squares estimator in the regular stochastic design convex regression model in dimensions 2 and 3. Besides the conjecture itself, the methods used in this section might be of independent interest. We will define a family of "localizing" functions that can be used to analyze the local properties of the least squares estimator. To the best of our knowledge, this thesis represents the first attempt to achieve this in a multidimensional scenario.

In change-point regression problems one tries to estimate a jump-discontinuity in an otherwise smooth function given a finite random sample. Change-point estimators tend to have the following characteristics: they converge at rate n^{-1}; their asymptotic theory is related to two-sided compound processes rather than to Gaussian processes; their limiting distributions have too many nuisance parameters, some of them living in infinite-dimensional spaces; despite being M-estimators, their asymptotic law cannot be deduced from the classical argmax continuous mapping theorem; the classical bootstrap yields inconsistent confidence intervals (for the concept of consistent bootstrap procedures, see Section 5.2.1). The second part of this document will illustrate all these properties of change-point problems and show ways to deal with them.

As mentioned in the previous paragraph, one of the peculiarities of change-point estimators is that they can usually be cast as M-estimators, but the traditional argmax continuous mapping theorem (see Theorem 3.2.2 in page 286 of van der Vaart and Wellner (1996)) cannot be used to derive their limiting laws. This happens because, in the limit, they are maximizers of two-sided compound Poisson processes which have multiple maximizers, almost surely. To remedy this situation, we force change-point estimators to be the smallest maximizers of their respective objective functions and then prove a version of the continuous mapping theorem pertinent to the situation. We carry this task in Chapter 4 and provide some examples in which the theorem can be applied.

Other relevant properties of change-point problems are that their limiting distributions depend on many nuisance parameters and that the classical bootstrap yields inconsistent confidence intervals. As the classical bootstrap is one of the most popular inferential techniques that avoid dealing with nuisance parameters, we carefully analyze this situation in Chapter 5. The failure of the classical bootstrap in nonstandard problems has been documented in several instances. For example, Bose and Chatterjee (2001), Abrevaya and Huang (2005) and Sen et al. (2010) have documented failure of the classical bootstrap in nonstandard, M-estimation problems (the former) and cube-root asymptotic problems (the latter 2). In Section 5.4 we will argue that the two most common bootstrap methods used in regression problems are inconsistent in the simplest change-point regression problem. Subsequently, two consistent methods for this problem will be provided in Section 5.5. While doing this, we will prove a consistency theorem for triangular arrays of random variables that might be of independent interest (see Proposition 5.3.3).

There are two main contributions of the analyses carried out in Chapters 4 and 5. On the one hand, the continuous mapping theorem of Chapter 4 is a convergence result that can be applied to many situations involving estimation of jump-discontinuities (see Li and Ling (2012) for an application in the context of threshold autoregressive models). On the other hand, the analysis of the consistency of the bootstrap schemes in Chapter 5 illustrates that the classical bootstrap cannot be trusted in change-point problems (as it fails in the simplest of such problems), but that smoothed bootstrap schemes are a consistent, easy-to-implement alternative. In addition, this work provides another instance of the inconsistency of the classical bootstrap in a nonstandard situation.

Part I

Convex Regression

Chapter 2

Multivariate Convex Regression

2.1 Least squares estimation of a multivariate convex regression function

Consider a closed, convex set $\mathfrak{X} \subset \mathbb{R}^d$, for $d \geq 1$, with nonempty interior and a regression model of the form

$$Y = \phi(X) + \epsilon \tag{2.1}$$

where X is a $\mathfrak{X}$-valued random vector, ϵ is a random variable with $\mathbf{E}(\epsilon\,|X) = 0$, and $\phi : \mathbb{R}^d \to \mathbb{R}$ is an unknown *convex* function. Given independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ from such a model, we wish to estimate ϕ by the method of least squares, i.e., by finding a convex function $\hat{\phi}_n$ which minimizes the discrete $\mathcal{L}_2$ norm

$$\left(\sum_{k=1}^{n} |Y_k - \psi(X_k)|^2 \right)^{\frac{1}{2}}$$

among all convex functions ψ defined on the convex hull of $X_1, \ldots, X_n$. In this paper we characterize the least squares estimator, provide means for its computation, study its finite sample properties and prove its consistency.

The problem just described is a nonparametric regression problem with known shape restriction (convexity). Such problems have a long history in the statistical literature with seminal papers like Brunk (1955), Grenander (1956) and Hildreth (1954) written more than 50 years ago, albeit in simpler settings. The former two papers deal with the estimation of monotone functions while the latter discusses least squares estimation of a concave function whose domain is a subset of the real line. Since then, many results on different nonparametric shape restricted regression problems have been published. For instance, Brunk (1970) and, more recently, Zhang (2002) have enriched the literature concerning isotonic regression. In the particular case of convex regression, Hanson and Pledger (1976) proved the consistency of the least squares estimator introduced in Hildreth (1954). Some years later, Mammen (1991) and Groeneboom et al. (2001) derived, respectively, the rate of convergence and asymptotic distribution of this estimator. Some alternative methods of estimation that combine shape restrictions with smoothness assumptions have also been proposed for the one-dimensional case; see, for example, Birke and Dette (2006) where a kernel-based estimator is defined and its asymptotic distribution derived.

Although the asymptotic theory of the one-dimensional convex regression problem is well understood, not much has been done in the multidimensional scenario. The absence of a natural order structure in $\mathbb{R}^d$, for $d > 1$, poses a natural impediment in such extensions. A convex function on the real line can be characterized as an absolutely continuous function with increasing first derivative (see, for instance, Folland (1999), Exercise 42.b, page 109). This characterization plays a key role in the computation and asymptotic theory of the least squares estimator in the one-dimensional case. By contrast, analogous results for convex functions of several variables involve more complicated characterizations using either second-order conditions (as in Dudley (1977), Theorem 3.1, page 163) or cyclical monotonicity (as in Rockafellar (1970), Theorems 24.8 and 24.9, pages 238-239). Interesting differences between convex functions on $\mathbb{R}$ and convex functions on $\mathbb{R}^d$ are given in Johansen (1974) and Bronšteĭn (1978).

Recently there has been considerable interest in shape restricted function estimation in multidimension. In the density estimation context, Cule et al. (2010) deal with the computation of the nonparametric maximum likelihood estimator of a multidimensional log-concave density, while Cule and Samworth (2010), Schuhmacher et al. (2009) and Schuhmacher and Dümbgen (2010) discuss its consistency and related issues. Seregin and Wellner (2009) study the computation and consistency of the maximum likelihood estimator of convex-transformed densities. This paper focuses on estimating a regression function which is known to be convex. To the best of our knowledge this is the first attempt to systematically study the characterization, computation, and consistency of the least squares estimator of a convex regression function with multidimensional covariates in a *completely nonparametric* setting.

In the field of econometrics some work has been done on this multidimensional problem in less general contexts and with more stringent assumptions. Estimation of concave and/or componentwise nondecreasing functions has been treated, for instance, in Banker and Maindiratta (1992), Matzkin (1991), Matzkin (1993), Beresteanu (2007) and Allon et al. (2007). The first two papers define maximum likelihood estimators in semiparametric settings. The estimators in Matzkin (1991) and Banker and Maindiratta (1992) are shown to be consistent in Matzkin (1991) and Maindiratta and Sarath (1997), respectively. A maximum likelihood estimator and a sieved least squares estimator have been defined and techniques for their computation have been provided in Allon et al. (2007) and Beresteanu (2007), respectively.

The method of least squares has been applied to multidimensional concave regression in Kuosmanen (2008). We take this work as our starting point. In agreement with the techniques used there, we define a least squares estimator which can be computed by solving a quadratic program. We argue that this estimator can be evaluated at a single point by finding the solution to a linear program. We then show that, under some mild regularity conditions, our estimator can be used to consistently estimate both, the convex function and its subdifferentials.

Our work goes beyond those mentioned above in the following ways: Our method does not require any tuning parameter(s), which is a major drawback for most nonparametric regression methods, such as kernel-based procedures. The choice of the tuning parameter(s) is especially problematic in higher dimensions, e.g., kernel based methods would require the choice of a $d \times d$ matrix of bandwidths. The sets of assumptions that most authors have used to study the estimation of a multidimensional convex regression function are more restrictive and of a different nature than the ones in this paper. As opposed to the maximum likelihood approach used in Banker and Maindiratta (1992), Matzkin (1991), Allon et al. (2007) and Maindiratta and Sarath (1997), we prove the consistency of the estimator keeping the distribution of the errors *unspecified*; e.g., in the i.i.d. case we only assume that the errors have zero expectation and finite second moment. The estimators in Beresteanu (2007) are sieved least squares estimators and assume that the observed values of the predictors lie on equidistant grids of rectangular domains. By contrast, our estimators are unsieved and our assumptions on the spatial arrangement of the predictor values are much more relaxed. In fact, we prove the consistency of the least squares estimator under both fixed and stochastic design settings; we also allow for heteroscedastic errors. In addition, we show that the least squares estimator can also be used to approximate the gradients and subdifferentials of the underlying convex function.

It is hard to overstate the importance of convex functions in applied mathematics. For instance, optimization problems with convex objective functions over convex sets appear in many applications. Thus, the question of accurately estimating a convex regression function is indeed interesting from a theoretical perspective. However, it turns out that convex regression is important for numerous reasons besides statistical curiosity. Convexity also appears in many applied sciences. One such field of application is microeconomic theory. Production functions are often supposed to be concave and componentwise nondecreasing. In this context, concavity reflects decreasing marginal returns. Concavity also plays a role in the theory of rational choice since it is a common assumption for utility functions, on which it represents decreasing marginal utility. The interested reader can see Hildreth (1954), Varian (1982a) or Varian (1982b) for more information regarding the importance of concavity/convexity in economic theory.

This chapter is organized as follows. In Section 2.2 we discuss the estimation procedure, characterize the estimator and show how it can be computed by solving a positive semidefinite quadratic program and a linear program. Section 2.3 starts with a description of the deterministic and stochastic design regression schemes. The statement and proof of our main results are also included in Section 2.3. In Section 2.4 we provide the proofs of the technical lemmas used to prove the main theorem. The appendix contains some auxiliary results from convex analysis and linear algebra that might be of independent interest.

2.2 Characterization and finite sample properties

We start with some notation. For convenience, we will regard elements of the Euclidian space $\mathbb{R}^m$ as column vectors and denote their components with upper indices, i.e, any $z \in \mathbb{R}^m$ will be denoted as $z = (z^1, z^2, \ldots, z^m)$. The symbol $\overline{\mathbb{R}}$ will stand for the extended real line. Additionally, for any set $A \subset \mathbb{R}^d$ we will denoted as $Conv\,(A)$ its convex hull and we'll write $Conv\,(X_1, \ldots, X_n)$ instead of $Conv\,(\{X_1, \ldots, X_n\})$. Finally, we will use $\langle \cdot, \cdot \rangle$ and $|\cdot|$ to denote the standard inner product and norm in Euclidian spaces, respectively.

For $\mathcal{X} = \{X_1, \ldots, X_n\} \subset \mathfrak{X} \subset \mathbb{R}^d$, consider the set $\mathcal{K}_{\mathcal{X}}$ of all vectors $z = (z^1, \ldots, z^n)' \in \mathbb{R}^n$ for which there is a convex function $\psi : \mathfrak{X} \to \mathbb{R}$ such that $\psi(X_j) = z^j$ for all $j = 1, \ldots, n$. Then, a necessary and sufficient condition for a convex function ψ to minimize the sum of squared errors is that $\psi(X_j) = Z_n^j$ for $j = 1, \ldots, n$, where

$$Z_n = \underset{z \in \mathcal{K}_{\mathcal{X}}}{\operatorname{argmin}} \left\{ \sum_{k=1}^{n} \left| Y_k - z^k \right|^2 \right\}. \tag{2.2}$$

The computation of the vector Z_n is crucial for the estimation procedure. We will show that such a vector exists and is unique. However, it should be noted that there are many convex functions ψ satisfying $\psi(X_j) = Z_n^j$ for all $j = 1, \ldots, n$. Although any of these functions can play the role of the least squares estimator, there is one such function which is easily evaluated in $Conv\,(X_1, \ldots, X_n)$. For computational convenience, we will define our least squares estimator $\hat{\phi}_n$ to be precisely this function and describe it explicitly in (2.7) and the subsequent discussion.

In what follows we show that both, the vector Z_n and the least squares estimator $\hat{\phi}_n$ are well-defined for any n data points $(X_1, Y_1), \ldots, (X_n, Y_n)$. We will also provide two characterizations of the set $\mathcal{K}_{\mathcal{X}}$ and show that the vector Z_n can be computed by solving a positive semidefinite quadratic program. Finally, we will prove that for any $x \in Conv\,(X_1, \ldots, X_n)$ one can obtain $\hat{\phi}_n(x)$ by solving a linear program.

2.2.1 Existence and uniqueness

We start with two characterizations of the set $\mathcal{K}_{\mathcal{X}}$. The developments here are similar to those in Allon et al. (2007) and Kuosmanen (2008).

Lemma 2.2.1 (Primal Characterization) *Let $z = (z^1, \ldots, z^n) \in \mathbb{R}^n$. Then, $z \in \mathcal{K}_{\mathcal{X}}$ if and only if for every $j = 1, \ldots, n$, the following holds:*

$$z^j = \inf \left\{ \sum_{k=1}^{n} \theta^k z^k : \sum_{k=1}^{n} \theta^k = 1,\ \sum_{k=1}^{n} \theta^k X_k = X_j,\ \theta \geq 0,\ \theta \in \mathbb{R}^n \right\}, \tag{2.3}$$

where the inequality $\theta \geq 0$ holds componentwise.

Proof: Define the function $g : \mathbb{R}^d \to \overline{\mathbb{R}}$ by

$$g(x) = \inf \left\{ \sum_{k=1}^{n} \theta^k z^k : \sum_{k=1}^{n} \theta^k = 1,\ \sum_{k=1}^{n} \theta^k X_k = x,\ \theta \geq 0,\ \theta \in \mathbb{R}^n \right\} \tag{2.4}$$

where we use the convention that $\inf(\emptyset) = +\infty$. By Lemma A.0.6 in the Appendix, g is convex and finite on the X_j's. Hence, if z^j satisfies (2.3) then $z^j = g(X_j)$ for every $j = 1, \ldots, n$ and it follows that $z \in \mathcal{K}_{\mathcal{X}}$.

Conversely, assume that $z \in \mathcal{K}_{\mathcal{X}}$ and $g(X_j) \neq z^j$ for some j. Note that $g(X_k) \leq z^k$ for any k from the definition of g. Thus, we may suppose that $g(X_j) < z^j$. As $z \in \mathcal{K}_{\mathcal{X}}$, there is a convex function ψ such that $\psi(X_k) = z^k$ for all $k = 1, \ldots, n$. Then, from the definition of $g(X_j)$ there exist $\theta_0 \in \mathbb{R}^n$ with $\theta_0 \geq 0$ and $\theta_0^1 + \ldots + \theta_0^n = 1$ such that $\theta_0^1 X_1 + \ldots + \theta_0^n X_n = X_j$ and

$$\sum_{k=1}^{n} \theta_0^k \psi(X_k) = \sum_{k=1}^{n} \theta_0^k z^k < z^j = \psi(X_j) = \psi\left(\sum_{k=1}^{n} \theta_0^k X_k\right),$$

which leads to a contradiction because ψ is convex. $\square$

We now provide an alternative characterization of the set $\mathcal{K}_{\mathcal{X}}$ based on the dual problem to the linear program used in Lemma 2.2.1.

Lemma 2.2.2 (Dual Characterization) *Let $z \in \mathbb{R}^n$. Then, $z \in \mathcal{K}_{\mathcal{X}}$ if and only if for any $j = 1, \ldots, n$ we have*

$$z^j = \sup\left\{\langle \xi, X_j\rangle + \eta : \langle \xi, X_k\rangle + \eta \le z^k \ \ \forall\ k = 1, \ldots, n,\ \xi \in \mathbb{R}^d,\ \eta \in \mathbb{R}\right\}. \quad (2.5)$$

Moreover, $z \in \mathcal{K}_{\mathcal{X}}$ if and only if there exist vectors $\xi_1, \ldots, \xi_n \in \mathbb{R}^d$ such that

$$\langle \xi_j, X_k - X_j\rangle \le z^k - z^j \ \ \forall\ k, j \in \{1, \ldots, n\}. \quad (2.6)$$

Proof: According to the primal characterization, $z \in \mathcal{K}_{\mathcal{X}}$ if and only if the linear programs defined by (2.3) have the z^j's as optimal values. The linear programs in (2.5) are the dual problems to those in (2.3). Then, the duality theorem for linear programs (see Luenberger (1984), page 89) implies that the z^j's have to be the corresponding optimal values to the programs in (2.5).

To prove the second assertion let us first assume that $z \in \mathcal{K}_{\mathcal{X}}$. For each $j \in \{1, \ldots, n\}$ take any solution (ξ_j, η_j) to (2.5). Then by (2.5), $\eta_j = z^j - \langle \xi_j, X_j\rangle$ and the inequalities in (2.6) follow immediately because we must have $\langle \xi_j, X_k\rangle + \eta_j \le z^k$ for any $k \in \{1, \ldots, n\}$. Conversely, take $z \in \mathbb{R}^n$ and assume that there are $\xi_1, \ldots, \xi_n \in \mathbb{R}^d$ satisfying (2.6). Take any $j \in \{1, \ldots, n\}$, $\eta_j = z^j - \langle \xi_j, X_j\rangle$ and θ to be the vector in $\mathbb{R}^n$ with components $\theta^k = \delta_{kj}$, where δ_{kj} is the Kronecker δ. It follows that $\langle \xi_j, X_k\rangle + \eta_j \le z^k\ \forall\ k = 1, \ldots, n$ so (ξ_j, η_j) is feasible for the linear program in (2.5). In addition, θ is feasible for the linear program in (2.3) so the weak duality principle of linear programming (see Luenberger (1984), Lemma 1, page 89) implies that $\langle \xi, X_j\rangle + \eta \le z^j$ for any pair (ξ, η) which is feasible for the problem in the right-hand side of

(2.5). We thus have that z^j is an upper bound attained by the feasible pair (ξ_j, η_j) and hence (2.5) holds for all $j = 1, \ldots, n$. $\square$

Both, the primal and dual characterizations are useful for our purposes. The primal plays a key role in proving the existence and uniqueness of the least squares estimator. The dual is crucial for its computation.

Lemma 2.2.3 *The set $\mathcal{K}_{\mathcal{X}}$ is a closed, convex cone in $\mathbb{R}^n$ and the vector Z_n satisfying (2.2) is uniquely defined.*

Proof: That $\mathcal{K}_{\mathcal{X}}$ is a convex cone follows trivially from the definition of the set. Now, if $z \notin \mathcal{K}_{\mathcal{X}}$, then there is $j \in \{1, \ldots, n\}$ for which $z^j > g(X_j)$ with the function g defined as in (2.4). Thus, there is $\theta_0 \in \mathbb{R}^n$ with $\theta_0 \geq 0$ and $\theta_0^1 + \ldots + \theta_0^n = 1$ such that $\theta_0^1 X_1 + \ldots + \theta_0^n X_n = X_j$ and $\sum_{k=1}^n \theta_0^k z^k < z^j$. Setting $\delta = \frac{1}{2}\left(z^j - \sum_{k=1}^n \theta_0^k z^k\right)$ it is easily seen that for all $\zeta \in \prod_{k=1}^n (z^k - \delta, z^k + \delta)$ we still have $\sum_{k=1}^n \theta_0^k \zeta^k < \zeta^j$ and thus $\zeta \notin \mathcal{K}_{\mathcal{X}}$. Therefore we have shown that for any $z \notin \mathcal{K}_{\mathcal{X}}$ there is a neighborhood U of z with $U \subset \mathbb{R}^n \setminus \mathcal{K}_{\mathcal{X}}$. Therefore, $\mathcal{K}_{\mathcal{X}}$ is closed and the vector Z_n is uniquely determined as the projection of $(Y_1, \ldots, Y_n) \in \mathbb{R}^n$ onto the closed convex set $\mathcal{K}_{\mathcal{X}}$ (see Conway (1985), Theorem 2.5, page 9). $\square$

We are now in a position to define the least squares estimator. Given observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ from model (2.1), we take the nonparametric least squares estimator to be the function $\hat{\phi}_n : \mathbb{R}^d \to \mathbb{R}$ defined by

$$\hat{\phi}_n(x) = \inf\left\{\sum_{k=1}^n \theta^k Z_n^k : \sum_{k=1}^n \theta^k = 1,\ \sum_{k=1}^n \theta^k X_k = x,\ \theta \geq 0,\ \theta \in \mathbb{R}^n\right\} \quad (2.7)$$

for any $x \in \mathbb{R}^d$. Here we are taking the convention that $\inf(\emptyset) = +\infty$. This function is well-defined because the vector Z_n exists and is unique for the sample. The estimator is, in fact, a polyhedral convex function (i.e., a convex function whose epigraph is a polyhedral; see Rockafellar (1970), page 172) and satisfies, as a consequence of Lemma A.0.6,

$$\hat{\phi}_n(x) = \sup_{\psi \in \mathcal{K}_{\mathcal{X},Z_n}} \{\psi(x)\},$$

where $\mathcal{K}_{\mathcal{X},Z_n}$ is the collection of all convex functions $\psi : \mathbb{R}^d \to \mathbb{R}$ such that $\psi(X_j) \leq Z_n^j$ for all $j = 1, \ldots, n$. Thus, $\hat{\phi}_n$ is the largest convex function that never exceeds the Z_n^j's. It is immediate that $\hat{\phi}_n$ is indeed a convex function (as the supremum of any family of convex functions is itself convex). The primal characterization of the set $\mathcal{K}_{\mathcal{X}}$ implies that $\hat{\phi}_n(X_j) = Z_n^j$ for all $j = 1, \ldots, n$.

2.2.2 Finite sample properties

In the following lemma we state some of the most important finite sample properties of the least squares estimator defined by (2.7).

Lemma 2.2.4 *Let $\hat{\phi}_n$ be the least squares estimator obtained from the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Then,*

(i) $\sum_{k=1}^{n} (\psi(X_k) - \hat{\phi}_n(X_k))(Y_k - \hat{\phi}_n(X_k)) \leq 0$ *for any convex function ψ which is finite on Conv $(X_1, \ldots, X_n)$;*

(ii) $\sum_{k=1}^{n} \hat{\phi}_n(X_k)(Y_k - \hat{\phi}_n(X_k)) = 0$*;*

(iii) $\sum_{k=1}^{n} Y_k = \sum_{k=1}^{n} \hat{\phi}_n(X_k)$*;*

(iv) the set on which $\hat{\phi}_n < \infty$ is Conv $(X_1, \ldots, X_n)$;

(v) for any $x \in \mathbb{R}^d$ the map $(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \mapsto \hat{\phi}_n(x)$ is a Borel-measurable function from $\mathbb{R}^{n(d+1)}$ into $\mathbb{R}$.

Proof: Property (i) follows from Moreau's decomposition theorem, which can be stated as:

Consider a closed convex set $\mathcal{C}$ on a Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. Then, for any $x \in \mathcal{H}$ there is only one vector $x_{\mathcal{C}} \in \mathcal{C}$ satisfying $\|x - x_{\mathcal{C}}\| = argmin_{\xi \in \mathcal{C}}\{\|x - \xi\|\}$. The vector $x_{\mathcal{C}}$ is characterized by being the only element of $\mathcal{C}$ for which the inequality $\langle \xi - x_{\mathcal{C}}, x - x_{\mathcal{C}} \rangle \leq 0$ holds for every $\xi \in \mathcal{C}$ (see Moreau (1962) or Song and Zhengjun (2004)).

Taking ψ to be $\kappa\hat{\phi}_n$ and letting κ vary through $(0, \infty)$ gives (ii) from (i). Similarly, (iii) follows from (i) by letting ψ to be $\hat{\phi}_n \pm 1$. Property (iv) is obvious from the definition of $\hat{\phi}_n$.

To see why (v) holds, we first argue that the map $(X_1, \dots, X_n, Y_1, \dots, Y_n) \mapsto Z_n$ is measurable. This follows from the fact that Z_n is the solution to a convex quadratic program and thus can be found as a limit of sequences whose elements come from arithmetic operations with $(X_1, \dots, X_n, Y_1, \dots, Y_n)$. Examples of such sequences are the ones produced by active set methods, e.g, see Boland (1997); or by interior-point methods (see Kapoor and Vaidya (1986) or Mehrotra and Sun (1990)). The measurability of $\hat{\phi}_n(x)$ follows from a similar argument, since it is the optimal value of a linear program whose solution can be obtained from arithmetic operations involving just $(X_1, \dots, X_n, Y_1, \dots, Y_n)$ and Z_n (e.g., via the well-known simplex method; see Nocedal and Wright (1999), page 372 or Luenberger (1984), page 30). $\square$

2.2.3 Computation of the estimator

Once the vector Z_n defined in (2.2) has been obtained, the evaluation of $\hat{\phi}_n$ at a single point x can be carried out by solving the linear program in (2.7). Thus, we need to find a way to compute Z_n. And here the dual characterization proves of vital importance, since it allows us to compute Z_n by solving a quadratic program.

Lemma 2.2.5 *Consider the positive semidefinite quadratic program*

$$\begin{array}{cc} \min & \sum_{k=1}^{n} |Y_k - z^k|^2 \\ \textit{subject to} & \langle \xi_k, X_j - X_k \rangle \leq z^j - z^k \quad \forall\ k, j = 1, \ldots, n \\ & \xi_1, \ldots, \xi_n \in \mathbb{R}^d, z \in \mathbb{R}^n. \end{array} \tag{2.8}$$

Then, this program has a unique solution Z_n in z, i.e., for any two solutions $(\xi_1, \ldots, \xi_n, z)$ and $(\tau_1, \ldots, \tau_n, \zeta)$ we have $z = \zeta = Z_n$. This solution Z_n is the only vector in $\mathbb{R}^n$ which satisfies (2.2).

Proof: From Lemma 2.2.2 if $(\xi_1, \ldots, \xi_n, z)$ belongs in the feasible set of this program, then $z \in \mathcal{K}_{\mathcal{X}}$. Moreover, for any $z \in \mathcal{K}_{\mathcal{X}}$ there are $\xi_1, \ldots, \xi_n \in \mathbb{R}^d$ such that $(\xi_1, \ldots, \xi_n, z)$ belongs to the feasible set of the quadratic program. Since the objective function only depends on z, solving the quadratic program is the same as getting the element of $\mathcal{K}_{\mathcal{X}}$ which is the closest to Y. This element is, of course, the uniquely defined Z_n satisfying (2.2). □

The quadratic program (2.8) is positive semidefinite. This implies certain computational complexities, but most modern nonlinear programming solvers can handle this type of optimization problems. Some examples of high-performance quadratic programming solvers are CPLEX, LINDO,
MOSEK and QPOPT. Here we present two simulated examples to illustrate the computation of the estimator when $d = 2$. The first one, depicted in Figure 2.1a corresponds to the case where $\phi(x) = |x|^2$. Figure 2.1b shows the convex function estimator when the regression function is the hyperplane $\phi(x) = -x^1 + x^2$. In both cases, $n = 256$ observations were used and the errors were assumed to be i.i.d. from the standard normal distribution. All the computations were carried out using the MOSEK optimization toolbox

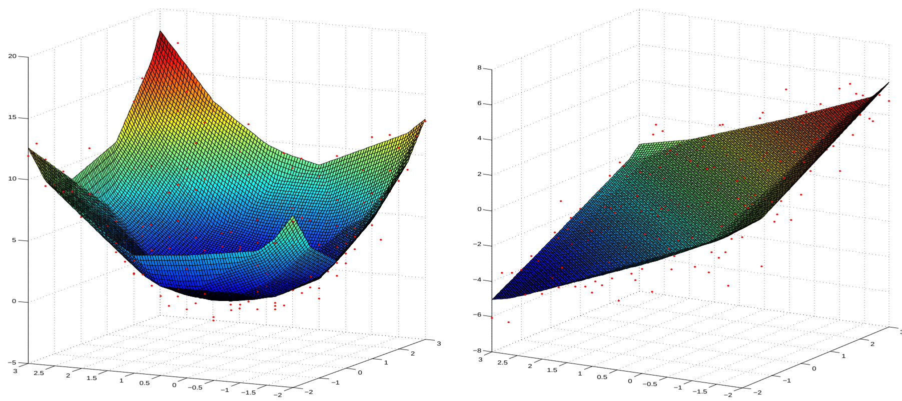

Figure 2.1: The scatter plot and nonparametric least squares estimator of the convex regression function when (a) $\phi(x) = |x|^2$ (left panel); (b) $\phi(x) = -x^1 + x^2$ (right panel).

for Matlab and the run time for each example was less than 2 minutes in a standard desktop PC. We refer the reader to Kuosmanen (2008) for additional numerical examples (although the examples there are for the estimation of concave, componentwise nondecreasing functions, the computational complexities are the same).

2.3 Consistency of the least squares estimator

The main goal of this paper is to show that in an appropriate setting the nonparametric least squares estimator $\hat{\phi}_n$ described above is consistent for estimating the convex function ϕ on the set $\mathfrak{X}$. In this context, we will prove the consistency of $\hat{\phi}_n$ in both, fixed and stochastic design regression settings.

Before proceeding any further we would like to introduce some notation. For any Borel set $\mathfrak{X} \subset \mathbb{R}^d$ we will denote by $\mathcal{B}_{\mathfrak{X}}$ the σ-algebra of Borel subsets of $\mathfrak{X}$. Given a sequence of events $(A_n)_{n=1}^{\infty}$ we will be using the notation

$[A_n \text{ i.o.}]$ and $[A_n \text{ a.a.}]$ to denote $\overline{\lim}\, A_n$ and $\underline{\lim}\, A_n$, respectively.

Now, consider a convex function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$. This function is said to be proper if $f(x) > -\infty$ for every $x \in \mathbb{R}^d$. The effective domain of f, denoted by $dom(f)$, is the set of points $x \in \mathbb{R}^d$ for which $f(x) < \infty$. The subdifferential of f at a point $x \in \mathbb{R}^d$ is the set $\partial f(x) \subset \mathbb{R}^d$ of all vectors ξ satisfying the inequality

$$\langle \xi, h \rangle \leq f(x+h) - f(x) \quad \forall\ h \in \mathbb{R}^d.$$

The elements of $\partial f(x)$ are called subgradients of f at x (see Rockafellar (1970)). For a set $A \subset \mathbb{R}^d$ we denote by A°, $\overline{A}$ and ∂A its interior, closure and boundary, respectively. We write $\text{Ext}(A) = \mathbb{R}^d \setminus \overline{A}$ for the exterior of the set A and $\text{diam}(A) := \sup_{x,y \in A} |x - y|$ for the diameter of A. We also use the sup-norm notation, i.e., for a function $g : \mathbb{R}^d \to \mathbb{R}$ we write $\|g\|_A = \sup_{x \in A} |g(x)|$.

To avoid measurability issues regarding some sets, specially those involving the random set-valued functions $\{\partial \hat{\phi}_n(x)\}_{x \in \mathfrak{X}^\circ}$, we will use the symbols $\mathbf{P}_*$ and $\mathbf{P}^*$ to denote inner and outer probabilities, respectively. We refer the reader to van der Vaart and Wellner (1996), pages 6-15, for the basic properties of inner and outer probabilities. In this context, a sequence of (not necessarily measurable) functions $(\Psi_n)_{n=1}^\infty$ from a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ into $\mathbb{R}$ is said to converge to a function Ψ *almost surely* (see van der Vaart and Wellner (1996), Definition 1.9.1-(iv), page 52), written $\Psi_n \xrightarrow{a.s.} \Psi$, if $\mathbf{P}_* (\Psi_n \to \Psi) = 1$. We will use the standard notation $\mathbf{P}(A)$ for the probabilities of all events A whose measurability can be easily inferred from the measurability of the random variables $\{\hat{\phi}_n(x)\}_{x \in \mathfrak{X}}$, established in Lemma 2.2.4.

Our main theorems hold for both, fixed and stochastic design schemes, and the proofs are very similar. They differ only in minor steps. Therefore, for the sake of simplicity, we will denote the observed values of the regressor variables always with the capital letters X_n. For any Borel set $\mathtt{X} \subset \mathbb{R}^d$, we write

$$N_n(\mathtt{X}) = \#\{1 \leq j \leq n : X_j \in \mathtt{X}\}.$$

The quantities X_n and $N_n(\mathtt{X})$ are non-random under the fixed design but random under the stochastic one.

2.3.1 Fixed Design

In a "fixed design" regression setting we assume that the regressor values are non-random and that all the uncertainty in the model comes from the response variable. We will now list a set of assumptions for this type of design. The one-dimensional case has been proven, under different regularity conditions, in Hanson and Pledger (1976).

(A1) We assume that we have a sequence $(X_n, Y_n)_{n=1}^{\infty}$ satisfying

$$Y_k = \phi(X_k) + \epsilon_k$$

where $(\epsilon_n)_{n=1}^{\infty}$ is an i.i.d. sequence with $\mathbf{E}(\epsilon_j) = 0$, $\mathbf{E}\left(\epsilon_j^2\right) = \sigma^2 < \infty$ and $\phi : \mathbb{R}^d \to \mathbb{R}$ is a proper convex function.

(A2) The non-random sequence $(X_n)_{n=1}^{\infty}$ is contained in a closed, convex set $\mathfrak{X} \subset \mathbb{R}^d$ with $\mathfrak{X}^\circ \neq \emptyset$ and $\mathfrak{X} \subset dom(\phi)$.

(A3) We assume the existence of a Borel measure ν on $\mathfrak{X}$ satisfying:

(i) $\{\mathtt{X} \in \mathcal{B}_{\mathfrak{X}} : \nu(\mathtt{X}) = 0\} = \{\mathtt{X} \in \mathcal{B}_{\mathfrak{X}} : \mathtt{X} \text{ has Lebesgue measure } 0\}$.

(ii) $\frac{1}{n} N_n(\mathtt{X}) \to \nu(\mathtt{X})$ for any open rectangle $\mathtt{X} \subset \mathfrak{X}^\circ$.

Condition (A1) may be replaced by the following:

(A4) We assume that we have a sequence $(X_n, Y_n)_{n=1}^{\infty}$ satisfying

$$Y_k = \phi(X_k) + \epsilon_k$$

where $\phi : \mathbb{R}^d \to \mathbb{R}$ is a proper convex function and $(\epsilon_n)_{n=1}^{\infty}$ is an independent sequence of random variables satisfying

(i) $\mathbf{E}(\epsilon_n) = 0 \ \forall\ n \in \mathbb{N}$ and $\underline{\lim} \frac{1}{n} \sum_{k=1}^{n} \mathbf{E}(|\epsilon_k|) > 0$.

(ii) $\sum_{n=1}^{\infty} \frac{\mathbf{Var}\left(\epsilon_n^2\right)}{n^2} < \infty$.

(iii) $\sup_{n\in\mathbb{N}}\{\mathbf{E}(\epsilon_n^2)\} < \infty$.

Under these conditions we define $\sigma^2 := \overline{\lim}_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \mathbf{E}\left(\epsilon_j^2\right)$.

The raison d'etre of condition (A4) is to allow the variance of the error terms to depend on the regressors. We make the distinction between (A1) and (A4) because in the case of i.i.d. errors it is enough to require a finite second moment to ensure consistency.

2.3.2 Stochastic Design

In this setting we assume that $(X_n, Y_n)_{n=1}^{\infty}$ is an i.i.d. sequence from some Borel probability measure μ on $\mathbb{R}^{d+1}$. Here we make the following assumptions on the measure μ:

(A5) There is a closed, convex set $\mathfrak{X} \subset \mathbb{R}^d$ with $\mathfrak{X}^{\circ} \neq \emptyset$ such that $\mu(\mathfrak{X} \times \mathbb{R}) = 1$. Also,

$$\int_{\mathfrak{X}\times\mathbb{R}} y^2 \mu(dx, dy) < \infty.$$

(A6) There is a proper convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ with $\mathfrak{X} \subset dom(\phi)$ such that whenever $(X, Y) \sim \mu$ we have $\mathbf{E}(Y - \phi(X)|X) = 0$ and $\mathbf{E}(|Y - \phi(X)|^2) = \sigma^2 < \infty$. Thus, ϕ is the regression function.

(A7) Denoting by $\nu(\cdot) = \mu((\cdot) \times \mathbb{R})$ the x-marginal of μ, we assume that

$$\{\mathtt{X} \in \mathcal{B}_{\mathfrak{X}} : \nu(\mathtt{X}) = 0\} = \{\mathtt{X} \in \mathcal{B}_{\mathfrak{X}} : \mathtt{X} \text{ has Lebesgue measure } 0\}.$$

We wish to point out some conclusions that one can draw from these assumptions. Consider the class of functions

$$\mathcal{K}_\nu := \left\{ \psi : \mathbb{R}^d \to \mathbb{R} \mid \psi \text{ is convex with } \int |\psi(x)|^2 \nu(dx) < \infty \right\}.$$

Then for any $\mathtt{X} \subset \mathfrak{X}$ the following holds

$$\int_{\mathtt{X}\times\mathbb{R}} \psi(x)(y - \phi(x))\mu(dx, dy) = 0 \ \ \forall\psi \ \in \mathcal{K}_\mu;$$

so we get that ϕ is in fact the element of $\mathcal{K}_\mu$ which is the closest to Y in the Hilbert space $\mathbb{L}^2(\mathtt{X} \times \mathbb{R}, \mathcal{B}_{\mathtt{X}\times\mathbb{R}}, \mu)$. This follows from Moreau's decomposition theorem (see the proof of Lemma 2.2.4).

Additionally, conditions {A5-A7} allow for stochastic dependency between the error variable $Y - \phi(X)$ and the regressor X. Although some level of dependency can be put to satisfy conditions {A2-A4}, the measure μ allows us to take into account some cases which wouldn't fit in the fixed design setting (even by conditioning on the regressors).

2.3.3 Main results

We can now state the two main results of this paper. The first result shows that assuming only the convexity of ϕ, the least squares estimator can be used to consistently estimate both ϕ and its subdifferentials $\partial\phi(x)$.

Theorem 2.3.1 *Under any of {A1-A3}, {A2-A4} or {A5-A7} we have,*

(i) $\mathbf{P}\left(\sup_{x\in X}\{|\hat{\phi}_n(x) - \phi(x)|\} \to 0 \ \ \textit{for any compact set } X \subset \mathfrak{X}^\circ \right) = 1.$

(ii) For every $x \in \mathfrak{X}^\circ$ and every $\xi \in \mathbb{R}^d$

$$\overline{\lim_{n\to\infty}} \lim_{h\downarrow 0} \frac{\hat{\phi}_n(x+h\xi) - \hat{\phi}_n(x)}{h} \leq \lim_{h\downarrow 0} \frac{\phi(x+h\xi)-\phi(x)}{h} \text{ almost surely.}$$

(iii) Denoting by $\boldsymbol{B}$ the unit ball (w.r.t. the Euclidian norm) we have

$$\mathbf{P}_* \left(\partial\hat{\phi}_n(x) \subset \partial\phi(x) + \epsilon\boldsymbol{B} \ \text{ a.a.}\right) = 1 \quad \forall\, \epsilon > 0,\ \forall\, x \in \mathfrak{X}^\circ.$$

(iv) If ϕ is differentiable at $x \in \mathfrak{X}^\circ$, then

$$\sup_{\xi \in \partial\hat{\phi}_n(x)} \{|\xi - \nabla\phi(x)|\} \xrightarrow{a.s.} 0.$$

Our second result states that assuming differentiability of ϕ on the entire $\mathfrak{X}^\circ$ allows us to use the subdifferentials of the least squares estimator to consistently estimate $\nabla\phi$ uniformly on compact subsets of $\mathfrak{X}^\circ$.

Theorem 2.3.2 *If ϕ is differentiable on $\mathfrak{X}^\circ$, then under any of $\{$A1-A3$\}$, $\{$A2-A4$\}$ or $\{$A5-A7$\}$ we have,*

$$\mathbf{P}_* \left(\sup_{\substack{\xi \in \partial\hat{\phi}_n(x) \\ x \in X}} \{|\xi - \nabla\phi(x)|\} \to 0 \ \text{ for any compact set } X \subset \mathfrak{X}^\circ \right) = 1.$$

2.3.4 Proof of the main results

Before embarking on the proofs, one must notice that there are some statements which hold true under any of {A1-A3}, {A2-A4} or {A5-A7}. We list the most important ones below, since they'll be used later.

- For any set $\mathtt{X} \subset \mathfrak{X}$ we have

$$\frac{N_n(\mathtt{X})}{n} \xrightarrow{a.s.} \nu(\mathtt{X}). \tag{2.9}$$

- The strong law of large numbers implies that for any Borel set $\mathtt{X} \subset \mathfrak{X}$ with positive Lebesgue measure we have

$$\frac{1}{N_n(\mathtt{X})} \sum_{\substack{X_k \in \mathtt{X} \\ 1 \leq k \leq n}} (Y_k - \phi(X_k)) \xrightarrow{a.s.} 0 \tag{2.10}$$

and also

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \sum_{1 \leq k \leq n} (Y_k - \phi(X_k))^2 = \sigma^2 \text{ a.s.} \tag{2.11}$$

We would like to point out that in the case of condition A4, A4-(iii) allows us to obtain (2.10) from an application of a version of the strong law of large number for uncorrelated random variables, as it appears in Chung (2001), page 108, Theorem 5.1.2. Similarly, condition A4-(ii) implies that we can apply a version the strong law of large numbers for independent random variables as in Williams (1991), Lemma 12.8, page 118 or in Folland (1999), Theorem 10.12, page 322 to obtain (2.11).

- For any Borel subset $\mathtt{X} \subset \mathfrak{X}$ with positive Lebesgue measure,

$$\#\{n \in \mathbb{N} : X_n \in \mathtt{X}\} \xrightarrow{a.s.} +\infty \tag{2.12}$$

Proof of Theorem 2.3.1. We will only make distinctions among the design schemes in the proof if we are using any property besides (2.9), (2.10), (2.11) or (2.12). For the sake of clarity, we divide the proof in steps.

Step I: We start by showing that for any set with positive Lebesgue measure there is a uniform band around the regression function (over that set) such that $\hat{\phi}_n$ comes within the band at least at one point for all but finitely many n's. This fact is stated in the following lemma (proved in Section 2.4.1).

Lemma 2.3.1 *For any set $\mathtt{X} \subset \mathfrak{X}$ with positive Lebesgue measure we have,*

$$\mathbf{P}\left(\inf_{x \in \mathtt{X}} \left\{|\hat{\phi}_n(x) - \phi(x)|\right\} \geq M \text{ i.o.}\right) = 0 \;\; \forall \; M > \frac{\sigma}{\sqrt{\nu(\mathtt{X})}}.$$

Step II: The idea is now to use the convexity of both, ϕ and $\hat{\phi}_n$, to show that the previous result in fact implies that the sup-norm of $\hat{\phi}_n$ is uniformly bounded on compact subsets of $\mathfrak{X}^\circ$. We achieve this goal in the following two lemmas (whose proofs are given in Sections 2.4.2 and 2.4.3 respectively).

Lemma 2.3.2 *Let $\mathtt{X} \subset \mathfrak{X}^\circ$ be compact with positive Lebesgue measure. Then, there is a positive real number $K_{\mathtt{X}}$ such that*

$$\mathbf{P}\left(\inf_{x \in \mathtt{X}} \{\hat{\phi}_n(x)\} < -K_{\mathtt{X}} \text{ i.o.}\right) = 0.$$

Lemma 2.3.3 *Let $\mathtt{X} \subset \mathfrak{X}^\circ$ be a compact set with positive Lebesgue measure. Then, there is $K_{\mathtt{X}} > 0$ such that*

$$\mathbf{P}\left(\sup_{x \in \mathtt{X}} \{\hat{\phi}_n(x)\} \geq K_{\mathtt{X}} \;\text{ i.o.}\right) = 0.$$

Step III: Convex functions are determined by their subdifferential mappings (see Rockafellar (1970), Theorem 24.9, page 239). Moreover, having a uniform upper bound $K_{\mathtt{X}}$ for the norms of all the subgradients over a compact region $\mathtt{X}$ imposes a Lipschitz continuity condition on the convex function over $\mathtt{X}$ (see Rockafellar (1970), Theorem 24.7, page 237); the Lipschitz constant being $K_{\mathtt{X}}$. For these reasons, it is important to have a uniform upper bound on the norms of the subgradients of $\hat{\phi}_n$ on compact regions. The following lemma (proved in Section 2.4.4) states that this can be achieved.

Lemma 2.3.4 *Let $\mathtt{X} \subset \mathfrak{X}^\circ$ be a compact set with positive Lebesgue measure. Then, there is $K_{\mathtt{X}} > 0$ such that*

$$\mathbf{P}^*\left(\sup_{\substack{\xi \in \partial\hat{\phi}_n(x) \\ x \in \mathtt{X}}} \{|\xi|\} > K_{\mathtt{X}} \text{ i.o.}\right) = 0.$$

Step IV: For the next results we need to introduce some further notation. We will denote by μ_n the empirical measure defined on $\mathbb{R}^{d+1}$ by the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. In agreement with van der Vaart and Wellner (1996), given a class of functions $\mathcal{G}$ on $D \subset \mathbb{R}^{d+1}$, a seminorm $\|\cdot\|$ on some space containing $\mathcal{G}$ and $\epsilon > 0$ we denote by $N(\epsilon, \mathcal{G}, \|\cdot\|)$ the ϵ covering number of $\mathcal{G}$ with respect to $\|\cdot\|$.

Although Lemmas 2.3.5 and 2.3.7 may seem unrelated to what has been done so far, they are crucial for the further developments. Lemma 3.5 (proved in Section 2.4.5) shows that the class of convex functions is not very complex in terms of entropy. Lemma 2.3.7 is a uniform version of the strong law of large numbers which proves vital in the proof of Lemma 2.3.8.

Lemma 2.3.5 *Let $\mathcal{X} \subset \mathfrak{X}^\circ$ be a compact rectangle with positive Lebesgue measure. For $K > 0$ consider the class $\mathcal{G}_{K,\mathcal{X}}$ of all functions of the form $\psi(X)(Y - \phi(X))\mathbf{1}_{\mathcal{X}}(X)$ where ψ ranges over the class $\mathcal{D}_{K,\mathcal{X}}$ of all proper convex functions which satisfy*

(a) $\|\psi\|_{\mathcal{X}} \leq K$;

(b) $\displaystyle\bigcup_{\substack{\xi \in \partial\psi(x) \\ x \in \mathcal{X}}} \{\xi\} \subset [-K, K]^d.$

Then, for any $\epsilon > 0$ we have

$$\varlimsup_{n\to\infty} N(\epsilon, \mathcal{G}_{K,\mathcal{X}}, \mathbb{L}_1(\mathcal{X} \times \mathbb{R}, \mu_n)) < \infty \quad \textit{almost surely,}$$

and there is a positive constant $A_\epsilon < \infty$, depending only on $(X_1, \ldots, X_n)$, K and $\mathcal{X}$, such that the covering numbers $N(\frac{\epsilon}{n}\sum_{j=1}^n |Y_j - \phi(X_j)|, \mathcal{G}_{K,\mathcal{X}}, \mathbb{L}_1(\mathcal{X} \times \mathbb{R}, \mu_n))$ are bounded above by A_ϵ, for all $n \in \mathbb{N}$, almost surely.

The proofs of Lemmas 2.3.7 and 2.3.8 (given in Sections 2.4.7 and 2.4.8 respectively) are the only parts in the whole proof where we must treat the different design schemes separately. To make the argument work, a small lemma (proved in Section 2.4.6) for the set of conditions {A2-A4} is required. We include it here for the sake of completeness and to point out the difference between the schemes.

Lemma 2.3.6 *Consider the set of conditions {A2-A4} and a subsequence* $(n_k)_{k=1}^{\infty}$ *such that*

$$\lim_{k\to\infty} \frac{1}{n_k} \sum_{j=1}^{n_k} \mathbf{E}\left(\epsilon_j^2\right) = \sigma^2.$$

Let $(\mathcal{X}_m)_{m=1}^{\infty}$ *be a an increasing sequence of compact subsets of* $\mathfrak{X}$ *satisfying* $\nu(X_m) \to 1$*. Then,*

$$\lim_{m\to\infty} \varliminf_{k\to\infty} \frac{1}{n_k} \sum_{\{1\le j\le n_k : X_j \in \mathcal{X}_m\}} \mathbf{E}\left(\epsilon_j^2\right) = \sigma^2.$$

We are now ready to state the key result on the uniform law of large numbers.

Lemma 2.3.7 *Consider the notation of Lemma 2.3.5 and let* $\mathcal{X} \subset \mathfrak{X}^{\circ}$ *be any finite union of compact rectangles with positive Lebesgue measure. Then,*

$$\sup_{\psi\in\mathcal{D}_{K,\mathcal{X}}} \left\{ \left| \frac{1}{n} \sum_{\{1\le j\le n : X_j\in\mathcal{X}\}} \psi(X_j)(Y_j - \phi(X_j)) \right| \right\} \xrightarrow{a.s.} 0.$$

Step V: With the aid of all the results proved up to this point, it is now possible to show that Lemma 2.3.1 is in fact true if we replace M by an arbitrarily small $\eta > 0$. The proof of the following lemma is given in Section 2.4.8.

Lemma 2.3.8 *Let* $\mathcal{X} \subset \mathfrak{X}^{\circ}$ *be any compact set with positive Lebesgue measure. Then,*

(i) $\mathbf{P}\left(\inf_{x\in\mathcal{X}}\{\phi(x) - \hat{\phi}_n(x)\} \ge \eta \ \text{ i.o.}\right) = 0 \quad \forall\, \eta > 0,$

(ii) $\mathbf{P}\left(\sup_{x\in X}\{\phi(x)-\hat{\phi}_n(x)\}\leq -\eta \;\text{ i.o.}\right)=0 \quad \forall\, \eta>0.$

Step VI: Combining the last lemma with the fact that we have a uniform bound on the norms of the subgradients on compacts, we can state and prove the consistency result on compacts. This is done in the next lemma (proof included in Section 2.4.9).

Lemma 2.3.9 *Let $X \subset \mathfrak{X}^\circ$ be a compact set with positive Lebesgue measure. Then,*

(i) $\mathbf{P}\left(\inf_{x\in X}\{\hat{\phi}_n(x)-\phi(x)\}<-\eta \;\text{ i.o.}\right)=0 \quad \forall\, \eta>0,$

(ii) $\mathbf{P}\left(\sup_{x\in X}\{\hat{\phi}_n(x)-\phi(x)\}>\eta \;\text{ i.o.}\right)=0 \quad \forall\, \eta>0,$

(iii) $\sup_{x\in X}\{|\hat{\phi}_n(x)-\phi(x)|\}\xrightarrow{a.s.} 0.$

Step VII: We can now complete the proof of Theorem 2.3.1. Consider the class $\mathfrak{C}$ of all open rectangles $\mathcal{R}$ such that $\overline{\mathcal{R}}\subset\mathfrak{X}^\circ$ and whose vertices have rational coordinates. Then, $\mathfrak{C}$ is countable and $\bigcup_{\mathcal{R}\in\mathfrak{C}}\mathcal{R}=\mathfrak{X}^\circ$. Observe that Lemmas 2.3.2 and 2.3.3 imply that for any finite union $A:=\mathcal{R}_1\cup\cdots\cup\mathcal{R}_m$ of open rectangles $\mathcal{R}_1,\ldots,\mathcal{R}_m\in\mathfrak{C}$ there is, with probability one, $n_0\in\mathbb{N}$ such that the sequence $(\hat{\phi}_n)_{n=n_0}^{\infty}$ is finite on $Conv\,(A)$. From Lemma 2.3.9 we know that the least squares estimator converges at all rational points in $\mathfrak{X}^\circ$ with probability one. Then, Theorem 10.8, page 90 of Rockafellar (1970) implies that (i) holds if $\mathfrak{X}^\circ$ is replaced by the convex hull of a finite union of rectangles belonging to $\mathfrak{C}$. Since there are countably many of such unions and any compact subset of $\mathfrak{X}^\circ$ is contained in one of those unions, we see that (i) holds. An application of Theorem 24.5, page 233 of Rockafellar (1970) on an open rectangle C containing x and satisfying $\overline{C}\subset\mathfrak{X}^\circ$ gives (ii) and (iii). Note that (iv) is a consequence of (iii). $\square$

Proof of Theorem 2.3.2. To prove the desired result we need the following lemma (whose proof is provided in Section 2.4.10) from convex analysis. The result is an extension of Theorem 25.7, page 248 of Rockafellar (1970), and might be of independent interest.

Lemma 2.3.10 *Let $\mathcal{C} \subset \mathbb{R}^d$ be an open, convex set and f a convex function which is finite and differentiable on $\mathcal{C}$. Consider a sequence of convex functions $(f_n)_{n=1}^{\infty}$ which are finite on $\mathcal{C}$ and such that $f_n \to f$ pointwise on $\mathcal{C}$. Then, if $\mathcal{X} \subset \mathcal{C}$ is any compact set,*

$$\sup_{\substack{x \in \mathcal{X} \\ \xi \in \partial f_n(x)}} \{|\xi - \nabla f(x)|\} \to 0.$$

Defining the class $\mathfrak{C}$ of open rectangles as in the proof of Theorem 2.3.1, one can use a similar argument to obtain Theorem 2.3.2 from an application of Theorem 2.3.1 and the previous lemma. $\square$

2.4 Proofs of auxiliary lemmas

Here we prove the lemmas involved in the proof of the main theorem. To prove these, we will need additional auxiliary results from matrix algebra and convex analysis, which may be of independent interest and are proved in the Appendix.

2.4.1 Proof of Lemma 2.3.1

We will first show that the event $\left[\inf_{x \in \mathtt{X}} \left\{\hat{\phi}_n(x) - \phi(x)\right\} \geq M \text{ i.o.}\right]$ has probability zero. Under this event, there is a subsequence $(n_k)_{k=1}^{\infty}$ such that $\inf_{x \in \mathtt{X}} \left\{\hat{\phi}_{n_k}(x) - \phi(x)\right\} \geq M \; \forall \, k \in \mathbb{N}$.

Then (2.10) implies that for this subsequence, with probability one, we have

$$\overline{\lim_{k\to\infty}} \frac{1}{N_{n_k}(\mathtt{X})} \sum_{X_j \in \mathtt{X}} \{Y_j - \hat{\phi}_{n_k}(X_j)\} \quad \le \quad -M. \tag{2.13}$$

On the other hand, it is seen (by solving the corresponding quadratic programming problems; see, e.g., Exercise 16.2, page 484 of Nocedal and Wright (1999)) that for any $\eta > 0$, $m \in \mathbb{N}$

$$\inf \left\{ \frac{1}{m} \sum_{1\le j\le m} |\xi^j|^2 \; : \frac{1}{m} \sum_{1\le j\le m} \xi^j \ge \eta, \; \xi \in \mathbb{R}^m \right\} \quad = \quad \eta^2, \tag{2.14}$$

$$\inf \left\{ \frac{1}{m} \sum_{1\le j\le m} |\xi^j|^2 \; : \frac{1}{m} \sum_{1\le j\le m} \xi^j \le -\eta, \; \xi \in \mathbb{R}^m \right\} \quad = \quad \eta^2. \tag{2.15}$$

For $0 < \delta < M$, using (2.15) with $\eta = M - \delta$ together with (2.12) and (2.13) we get that, with probability one, we must have

$$\underline{\lim_{k\to\infty}} \frac{1}{n_k} \sum_{j=1}^{n_k} (Y_j - \hat{\phi}_{n_k}(X_j))^2 \ge \nu(\mathtt{X})(M-\delta)^2.$$

Letting $\delta \to 0$ we actually get

$$\underline{\lim_{k\to\infty}} \frac{1}{n_k} \sum_{j=1}^{n_k} (Y_j - \hat{\phi}_{n_k}(X_j))^2 \ge \nu(\mathtt{X}) M^2 > \sigma^2 = \overline{\lim_{k\to\infty}} \frac{1}{n_k} \sum_{j=1}^{n_k} (Y_j - \phi(X_j))^2 \;\text{ a.s.}$$

which is impossible because $\hat{\phi}_{n_k}$ is the least squares estimator. Therefore,

$$\mathbf{P}\left(\inf_{x\in\mathtt{X}} \left\{ \hat{\phi}_n(x) - \phi(x) \right\} \ge M \text{ i.o.} \right) = 0.$$

A similar argument now using (2.14) gives

$$\mathbf{P}\left(\sup_{x\in\mathtt{X}} \left\{ \hat{\phi}_n(x) - \phi(x) \right\} \le -M \text{ i.o.} \right) = 0,$$

which completes the proof of the lemma. $\square$

Before we prove Lemmas 2.3.2 and 2.3.3, we need some additional results from matrix algebra. For convenience, we state them here, but postpone their proofs to Section B in the Appendix.

We first introduce some notation. We write $\mathbf{e}_j \in \mathbb{R}^d$ for the vector whose components are given by $\mathbf{e}_j^k = \delta_{jk}$, where δ_{jk} is the Kronecker δ. We also write $\mathbf{e} = \mathbf{e}_1 + \ldots + \mathbf{e}_d$ for the vector of ones in $\mathbb{R}^d$. For $\alpha \in \{-1,1\}^d$ we write

$$\mathcal{R}_\alpha = \left\{ \sum_{k=1}^{d} \theta^k \alpha^k \mathbf{e}_k : \theta \geq 0, \theta \in \mathbb{R}^d \right\}$$

for the orthant in the α direction. For any hyperplane $\mathcal{H}$ defined by the normal vector $\xi \in \mathbb{R}^d$ and the intercept $b \in \mathbb{R}$, we write $\mathcal{H} = \{x \in \mathbb{R}^d : \langle \xi, x \rangle = b\}$, $\mathcal{H}^+ = \{x \in \mathbb{R}^d : \langle \xi, x \rangle > b\}$ and $\mathcal{H}^- = \{x \in \mathbb{R}^d : \langle \xi, x \rangle < b\}$. For $r > 0$ and $x_0 \in \mathbb{R}^d$ we will write $B(x_0, r) = \{x \in \mathbb{R}^d : |x - x_0| < r\}$. We denote by $\mathbb{R}^{d \times d}$ the space of $d \times d$ matrices endowed with the topology defined by the $\| \cdot \|_2$ norm (where $\|A\|_2 = \sup_{|x| \leq 1}\{|Ax|\}$ and can be shown to be equal to the largest singular value of A; see Harville (2008)).

Lemma 2.4.1 *Let $r > 0$. There is a constant $R_r > 0$, depending only on r and d, such that for any $\rho_* \in (0, R_r)$ there are $\rho, \rho^* > 0$ with the property: for any $\alpha \in \{-1,1\}^d$ and any d-tuple of vectors $\beta = \{x_1, \ldots, x_d\} \subset \mathbb{R}^d$ such that $x_j \in B(\alpha^j r \boldsymbol{e}_j, \rho)$ $\forall$ $j = 1, \ldots, d$, there is a unique pair $(\xi_{\alpha,\beta}, b_{\alpha,\beta})$, with $\xi_{\alpha,\beta} \in \mathbb{R}^d$, $|\xi_{\alpha,\beta}| = 1$ and $b_{\alpha,\beta} > 0$ for which the following statements hold:*

(i) β *form a basis for* $\mathbb{R}^d$.

(ii) $x_1, \ldots, x_d \in \mathcal{H}_{\alpha,\beta} := \{x \in \mathbb{R}^d : \langle \xi_{\alpha,\beta}, x \rangle = b_{\alpha,\beta}\}$.

(iii) $\min\limits_{1 \leq j \leq d} \{|\xi^j_{\alpha,\beta}|\} > 0$.

(iv) $B(0, \rho_*) \subset \mathcal{H}^-_{\alpha,\beta}$.

(v) $\{x \in \mathbb{R}^d : |x| \geq \rho^*\} \cap \mathcal{R}_\alpha \subset \mathcal{H}^+_{\alpha,\beta}$.

(vi) $B(-\alpha^j r \boldsymbol{e}_j, \rho) \subset \{x \in \mathbb{R}^d : \langle \xi_{\alpha,\beta}, x \rangle < 0\}$ *for all* $j = 1, \ldots, d$.

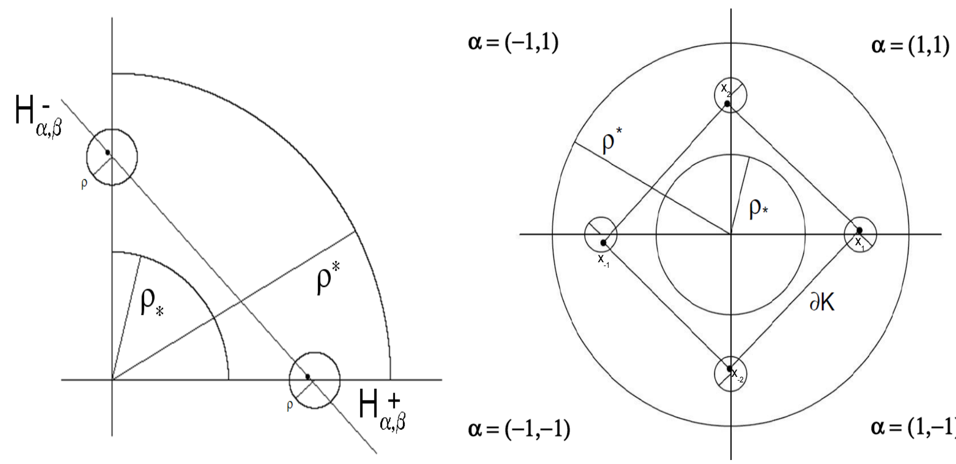

Figure 2.2: Explanatory diagram for (a) Lemma 2.4.1 (left panel); (b) Lemma 2.4.2 (right panel).

(vii) For any $w_1 \in B\left(0, \frac{\rho_*}{16\sqrt{d}}\right)$ *and* $w_2 \in B\left(\frac{3\rho_*}{8\sqrt{d}}\alpha, \frac{\rho_*}{8\sqrt{d}}\right)$ *we have*

$$\min_{1\le j\le d}\left\{\left(X_\beta^{-1}\left(w_1+t(w_2-w_1)\right)\right)^j\right\} > 0 \quad \forall\, t \ge 1$$

where $X_\beta = (x_1, \dots, x_d) \in \mathbb{R}^{d\times d}$ *is the matrix whose* j*'th column is* x_j*.*

Figure 2.2a illustrates the above lemma when $d = 2$ and $\alpha = (1,1)$. The lemma states that whatever points x_1 and x_2 are taken inside the circles of radius ρ around $\alpha^1 r\mathbf{e}_1$ and $\alpha^2 r\mathbf{e}_2$, respectively, $B(0,\rho_*)$ and $\{x \in \mathbb{R}^d : |x| \ge \rho^*\} \cap \mathcal{R}_\alpha$ are contained, respectively, in the half-spaces $\mathcal{H}^-_{\alpha,\beta}$ and $\mathcal{H}^+_{\alpha,\beta}$. Assertion (vii) of the lemma implies that all the points in the half line $\{w_1 + t(w_2 - w_1\}_{t\ge1}$ should have positive co-ordinates with respect to the basis β as they do with respect to the basis $\{\alpha^j\mathbf{e}_j\}_{j=1}^d$. We refer the reader to Section B.1 for a complete proof of Lemma 2.4.1.

We now state two other useful results, namely Lemma 2.4.2 and Lemma 2.4.3, but defer their proofs to Section B.2 and Section B.3 respectively.

Lemma 2.4.2 *Let* $r > 0$ *and consider the notation of Lemma 2.4.1 with the positive numbers* ρ*,* ρ_* *and* ρ^* *as defined there. Take 2d vectors* $\{x_{\pm1}, \dots, x_{\pm d}\}$

$\subset \mathbb{R}^d$ *such that* $x_{\pm j} \in B(\pm r\boldsymbol{e}_j, \rho)$ *and for* $\alpha \in \{-1,1\}^d$ *write* $\beta_\alpha = \{x_{\alpha^1 1}, x_{\alpha^2 2}, \ldots, x_{\alpha^d d}\}$*,* $\xi_\alpha = \xi_{\alpha,\beta_\alpha}$*,* $b_\alpha = b_{\alpha,\beta_\alpha}$ *and* $\mathcal{H}_\alpha = \mathcal{H}_{\alpha,\beta}$*, all in agreement with the setting of Lemma 2.4.1. Then, if* $K = Conv\,(x_{\pm 1}, \ldots, x_{\pm d})$ *we have:*

(i) $K = \bigcap_{\alpha \in \{-1,1\}^d} \{x \in \mathbb{R}^d : \langle \xi_\alpha, x \rangle \leq b_\alpha\}$.

(ii) $K^\circ = \bigcap_{\alpha \in \{-1,1\}^d} \{x \in \mathbb{R}^d : \langle \xi_\alpha, x \rangle < b_\alpha\}$.

(iii) $\partial K = \bigcup_{\alpha \in \{-1,1\}^d} Conv\,(x_{\alpha^1 1}, \ldots, x_{\alpha^d d})$.

(iv) $\partial K = \left(\bigcup_{\alpha \in \{-1,1\}^d} \{x \in \mathbb{R}^d : \langle \xi_\alpha, x \rangle = b_\alpha\}\right) \bigcap \left(\bigcap_{\alpha \in \{-1,1\}^d} \{x \in \mathbb{R}^d : \langle \xi_\alpha, x \rangle \leq b_\alpha\}\right)$.

(v) $B(0, \rho_*) \subset K^\circ$.

(vi) $\partial B(0, \rho^*) \subset \mathrm{Ext}(K)$.

Figure 2.2b illustrates Lemma 2.4.2 for the two-dimensional case. Intuitively, the idea is that as long as the points $x_{\pm 1}$ and $x_{\pm 2}$ belong to $B(\pm r\mathbf{e}_1, \rho)$ and $B(\pm r\mathbf{e}_2, \rho)$, respectively, we will have $B(0, \rho_*)$ and $\partial B(0, \rho^*)$ as subsets of K° and $\mathrm{Ext}(K)$, respectively.

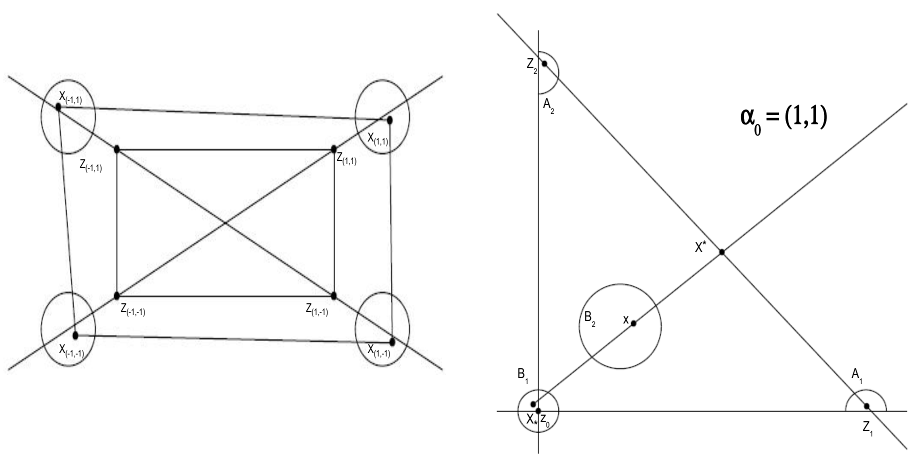

Figure 2.3: Explanatory diagram for (a) Lemma 2.4.3 (left panel); (b) Lemma 2.3.2 (right panel).

Lemma 2.4.3 *Let $[a,b] \subset \mathbb{R}^d$ be a compact rectangle and $r > 0$, with $r < \frac{1}{d-2}$ if $d \geq 3$. For each $\alpha \in \{-1,1\}^d$ write $z_\alpha = a + \sum_{j=1}^d \frac{1+\alpha^j}{2}(b^j - a^j)\,\boldsymbol{e}_j$ so that $\{z_\alpha\}_{\alpha\in\{-1,1\}^d}$ is the set of vertices of $[a,b]$. Then, there is $\rho > 0$ such that if $x_\alpha \in B(z_\alpha + r(z_\alpha - z_{-\alpha}), \rho)\ \forall\ \alpha \in \{-1,1\}^d$, then*

$$[a,b] \subset Conv\left(x_\alpha : \alpha \in \{-1,1\}^d\right)^\circ.$$

Figure 2.3a describes Lemma 2.4.3 in the two-dimensional case. As long as the points $x_{(\pm1,\pm1)}$ are chosen in the balls of radius ρ around $z_{(\pm1,\pm1)} + r(z_{(\pm1,\pm1)} - z_{(\mp1,\mp1)})$, $Conv\left(x_{(\pm1,\pm1)}\right)$ will contain $Conv\left(z_{(\pm1,\pm1)}\right)$.

2.4.2 Proof of Lemma 2.3.2

Since any compact subset of $\mathfrak{X}^\circ$ is contained in a finite union of compact rectangles, it is enough to prove the result when $\mathtt{X}$ is a compact rectangle $[a,b] \subset \mathfrak{X}^\circ$. Let $r = \frac{1}{4}\min_{1\leq k\leq d}\{b^k - a^k\}$ and choose $\rho \in (0, \frac{1}{4}r)$, $\rho^* > 0$ and $0 < \rho_* < \frac{1}{2}r$ such that the conclusions of Lemmas 2.4.1 and 2.4.2 hold for any $\alpha \in \{-1,1\}^d$ and any $\beta = (z_1, \ldots, z_d) \in \mathbb{R}^{d\times d}$ with $z_j \in B(\alpha^j r\mathbf{e}_j, \rho)$. Take $N \in \mathbb{N}$ such that

$$\frac{1}{N}\max_{1\leq k\leq d}\{b^k - a^k\} < \frac{1}{32d}\rho_* \tag{2.16}$$

and divide $\mathtt{X}$ into N^d rectangles all of which are geometrically identical to $\frac{1}{N}[0, b-a]$. Let $\mathcal{C}$ be any one of the rectangles in the grid and choose any vertex z_0 of $\mathcal{C}$ satisfying

$$z_0 = \underset{z\in\mathcal{C}}{\operatorname{argmax}}\left\{\max_{1\leq j\leq d}\left\{z^j - a^j, b^j - z^j\right\}\right\}.$$

Then, from the definition of z_0 and r, there is $\alpha_0 \in \{-1,1\}^d$ such that

$$B(z_0, r) \cap (z_0 + \mathcal{R}_{\alpha_0}) \subset \mathtt{X}.$$

Additionally, define

$$\begin{aligned}
B_1 &= B\left(z_0, \frac{\rho_*}{16\sqrt{d}}\right), \\
B_2 &= B\left(z_0 + \frac{3\rho_*}{8\sqrt{d}}\alpha_0, \frac{\rho_*}{8\sqrt{d}}\right), \\
A_j &= B(z_0 + \alpha_0^j r\mathbf{e}_j, \rho) \cap (z_0 + \mathcal{R}_{\alpha_0}) \quad \forall\ j = 1, \ldots, d, \\
A_{-j} &= B(z_0 - \alpha_0^j r\mathbf{e}_j, \rho) \quad \forall\ j = 1, \ldots, d.
\end{aligned}$$

Observe that all the sets in the previous display have positive Lebesgue measure and that the A_{-j}'s are not necessarily contained in $\mathtt{X}$. Let $M_1 = \|\phi\|_{\mathtt{X}}$, $M_0 > \frac{\sigma}{\sqrt{\min\{\nu(B_1),\nu(B_2),\nu(A_1),\ldots,\nu(A_d)\}}}$, $M = M_1 + M_0$ and $K_{\mathcal{C}} > 6M$. Also, notice that $\mathcal{C} \subset B_1$ because of (2.16). We will argue that

$$\mathbf{P}\left(\inf_{x\in\mathcal{C}}\{\hat{\phi}_n(x)\} \leq -K_{\mathcal{C}} \text{ i.o.}\right) = 0. \tag{2.17}$$

From Lemma 2.3.1, we know that

$$\mathbf{P}\left(\bigcap_{j=1}^{d}\left[\inf_{x\in A_j}\left\{\left|\hat{\phi}_n(x) - \phi(x)\right|\right\} < M_0 \text{ a.a.}\right]\right) = 1, \tag{2.18}$$

so there is, with probability one, $n_0 \in \mathbb{N}$ such that $\inf_{x\in A_j}\left\{\left|\hat{\phi}_n(x) - \phi(x)\right|\right\} < M_0$ for any $n \geq n_0$ and any $j = 1, \ldots, d$.

Assume that the event $\left[\inf_{x\in\mathcal{C}}\{\hat{\phi}_n(x)\} < -K_{\mathcal{C}} \text{ i.o.}\right]$ is true. Then, there is a subsequence n_k such that $\inf_{x\in\mathcal{C}}\{\hat{\phi}_{n_k}(x)\} < -K_{\mathcal{C}}$ for all $k \in \mathbb{N}$. Fix any $k \geq n_0$. We know that there is $X_* \in \mathcal{C} \subset B_1$ such that $\hat{\phi}_{n_k}(X_*) \leq -K_{\mathcal{C}}$. In addition, for $j = 1, \ldots, d$, there are $Z_{\alpha_0^j j} \in A_j$ such that $|\hat{\phi}_{n_k}(Z_{\alpha_0^j j}) - \phi(Z_{\alpha_0^j j})| < M_0$, which in turn implies $\hat{\phi}_{n_k}(Z_{\alpha_0^j j}) < M$. Pick any $Z_{-\alpha_0^j} \in A_{-j}$ and let $K = Conv\,(Z_{\pm 1}, \ldots, Z_{\pm d}) = z_0 + Conv\,(Z_{\pm 1} - z_0, \ldots, Z_{\pm d} - z_0)$.

Take any $x \in B_2$. We will show the existence of $X^* \in Conv\left(Z_{\alpha_0^1 1}, \ldots, Z_{\alpha_0^d d}\right)$ such that $x \in Conv\,(X_*, X^*)$, as shown in Figure 2.3b for the case $d = 2$. We will then show that the existence of such an X^* implies that

$$|\phi(x) - \hat{\phi}_{n_k}(x)| > M_0. \tag{2.19}$$

Consequently, since x is an arbitrary element of B_2 we will have

$$\left[\inf_{x\in\mathcal{C}}\{\hat{\phi}_n(x)\} \le -K_{\mathcal{C}} \text{ i.o.}\right] \cap \left(\bigcap_{j=1}^{d}\left[\inf_{x\in A_j}\left\{\left|\hat{\phi}_n(x)-\phi(x)\right|\right\} < M_0 \text{ a.a.}\right]\right)$$

$$\subset \left[\inf_{x\in B_2}\{|\phi(x)-\hat{\phi}_{n_k}(x)|\} \ge M_0 \text{ i.o.}\right].$$

But from Lemma 2.3.1, the event on the right is a null set. Taking (2.18) into account, we will see that (2.17) holds and then complete the argument by taking $K_{\mathtt{X}} = \max_{\mathcal{C}}\{K_{\mathcal{C}}\}$.

To show the existence of X^* consider the function $\psi : \mathbb{R} \to \mathbb{R}^d$ given by $\psi(t) = X_* + t(x - X_*)$. The function ψ is clearly continuous and satisfies $\psi(0) = X_*$ and $\psi(1) = x \in B_2 \subset K^\circ$. That $B_2 \subset K^\circ$ is a consequence of Lemma 2.4.1, (iv). The set K is bounded, so there is $T > 1$ such that $\psi(T) \in \text{Ext}(K) = \mathbb{R}^d \setminus \overline{K}$. The intermediate value theorem then implies that there is $t^* \in (1, T)$ such that $X^* := \psi(t^*) \in \partial K$. Observe that by Lemma 2.4.2 (iii) we have

$$\partial K = \bigcup_{\alpha\in\{-1,1\}^d} Conv\left(Z_{\alpha^1 1}, \dots, Z_{\alpha^d d}\right).$$

Lemma 2.4.1 (i) implies that $\{Z_{\alpha_0^1 1} - z_0, \dots, Z_{\alpha_0^d d} - z_0\}$ forms a basis of $\mathbb{R}^d$ so we can write $X^* - z_0 = \sum_{j=1}^d \theta^j (Z_{\alpha_0^j j} - z_0)$. Moreover, Lemma 2.4.1 (vii) implies that $\theta^j > 0$ for every $j = 1, \dots, d$ as $\theta = (\theta^1, \dots, \theta^d) = (Z_{\alpha_0^1 1} - z_0, \dots, Z_{\alpha_0^d d} - z_0)^{-1}(X^* - z_0)$. Here we apply Lemma 2.4.1 (vii) with $w_1 = X_* \in B_1$, $w_2 = x \in B_2$ and $t^* > 1$.

For $\alpha \in \{-1, 1\}^d$ consider the pair $(\xi_\alpha, b_\alpha) \in \mathbb{R}^d \times \mathbb{R}$ as defined in Lemma 2.4.2 for the set of vectors $\{Z_{\pm 1} - z_0, \dots, Z_{\pm d} - z_0\}$ (here we move the origin to z_0). Observe that Lemma 2.4.1 (ii) implies that $\langle \xi_{\alpha_0}, Z_{\alpha_0^j j} - z_0 \rangle = b_{\alpha_0}$ for all $j = 1, \dots, d$. Consequently, $\langle \xi_{\alpha_0}, X^* - z_0 \rangle = b_{\alpha_0} \sum_{j=1}^d \theta^j$, but since

$X^* \in \partial K$, Lemma 2.4.2 (iv) implies that $\langle \xi_{\alpha_0}, X^* - z_0 \rangle \leq b_{\alpha_0}$ and hence $\sum_{j=1}^{d} \theta^j \leq 1$. Additionally, for $\alpha \neq \alpha_0$ we can write $\langle \xi_\alpha, X^* - z_0 \rangle$ as

$$\sum_{j=1}^{d} \theta^j \langle \xi_\alpha, Z_{\alpha_0^j j} - z_0 \rangle = \sum_{\alpha^j = \alpha_0^j} \theta^j b_\alpha + \sum_{\alpha^j \neq \alpha_0^j} \theta^j \langle \xi_\alpha, Z_{\alpha_0^j j} - z_0 \rangle < b_\alpha \tag{2.20}$$

as $\langle \xi_\alpha, Z_{\alpha^j} - z_0 \rangle = b_\alpha$ (by Lemma 2.4.1 (ii)) and $\langle \xi_\alpha, Z_{-\alpha^j} - z_0 \rangle < 0$ (by Lemma 2.4.1 (vi)) for every $j = 1, \ldots, d$. Since $\langle \xi_\alpha, w - z_0 \rangle = b_\alpha$ for all $w \in Conv\,(Z_{\alpha^1 1}, \ldots, Z_{\alpha^d d})$ and all $\alpha \in \{-1.1\}^d$, (2.20) and the fact that $X^* \in \partial K$ imply that $X^* \in Conv\left(Z_{\alpha_0^1 1}, \ldots, Z_{\alpha_0^d d}\right)$. Hence $\hat{\phi}_n(X^*) \leq \sum_{j=1}^{d} \theta^j \hat{\phi}_{n_k}(Z_{\alpha_0^j j}) < M$. We therefore have

$$\hat{\phi}_{n_k}(X^*) < M \quad , \quad \hat{\phi}_{n_k}(X_*) < -K_{\mathcal{C}}, \tag{2.21}$$

$$X_* + \frac{1}{t^*}(X^* - X_*) \;=\; x. \tag{2.22}$$

Since $X_* \in B_1$ and $d \geq 1$ we have

$$|z_0 - X_*| < \frac{1}{8}\rho_*. \tag{2.23}$$

By using the triangle inequality we get the following bounds

$$\frac{1}{4}\rho_* < |z_0 - x| < \frac{1}{2}\rho_*. \tag{2.24}$$

And from Lemma 2.4.1 (iv) and the fact that $\langle \xi_{\alpha_0}, X^* \rangle = b_{\alpha_0}$ we also obtain

$$|z_0 - X^*| \geq \rho_*. \tag{2.25}$$

From (2.22) we know that $t^* = \frac{|X^* - X_*|}{|x - X_*|}$. Using the triangle inequality with (2.23), (2.24) and (2.25) one can find lower and upper bounds for $|X^* - X_*|$ (as $|X^* - X_*| \geq |X^* - z_0| - |z_0 - X_*|$) and $|x - X_*|$ (as $|x - X_*| \leq |x - z_0| + |z_0 - X_*|$), respectively, to obtain $t^* \geq \frac{7}{5}$. Then, (2.21) and (2.22) imply

$$\hat{\phi}_{n_k}(x) \leq \left(1 - \frac{1}{t^*}\right) \hat{\phi}_{n_k}(X_*) + \frac{1}{t^*} \hat{\phi}_{n_k}(X^*) \leq -\frac{2}{7} K_{\mathcal{C}} + \frac{5}{7} M < -M.$$

Consequently,

$$|\phi(x) - \hat{\phi}_{n_k}(x)| > M - M_1 = M_0.$$

This proves (2.19) and completes the proof. $\square$

2.4.3 Proof of Lemma 2.3.3

Assume without loss of generality that $\mathtt{X}$ is a compact rectangle. Let $\{z_\alpha : \alpha \in \{-1,1\}^d\}$ be the set of vertices of the rectangle. Then, there is $r \in (0,1)$ such that $B(z_\alpha, r) \subset \mathfrak{X}^\circ\ \forall\ \alpha \in \{-1,1\}^d$. Recall that from Lemma 2.4.3, there is $0 < \rho < \frac{1}{2}r$ such that for any $\{\eta_\alpha : \alpha \in \{-1,1\}^d\}$ if $\eta_\alpha \in B(z_\alpha + \frac{r}{2}(z_\alpha - z_{-\alpha}), \rho)$ then $\mathtt{X} \subset Conv\left(\eta_\alpha : \alpha \in \{-1,1\}^d\right)$.

Let $A_\alpha = B(z_\alpha + \frac{1}{2}r(z_\alpha - z_{-\alpha}), \frac{\rho}{2})$ and $M_0 > \frac{\sigma}{\sqrt{\min\{\nu(A_\alpha):\alpha\in\{-1,1\}^d\}}}$ and choose

$$M_1 = \sup_{x \in Conv\left(\bigcup_{\alpha\in\{-1,1\}^d} A_\alpha\right)} \{|\phi(x)|\}.$$

Take $K_{\mathtt{X}} > M_0 + M_1$. Since

$$\mathbf{P}\left(\bigcap_{\alpha\in\{-1,1\}^d}\left[\inf_{x\in A_\alpha}\{|\hat{\phi}_n(x) - \phi(x)|\} < M_0,\ \text{a.a.}\right]\right) = 1$$

by Lemma 2.3.1, there is, with probability one, $n_0 \in \mathbb{N}$ such that for any $n \geq n_0$ we can find $\eta_\alpha \in A_\alpha$, $\alpha \in \{-1,1\}^d$, such that $|\hat{\phi}_n(\eta_\alpha) - \phi(\eta_\alpha)| < M_0$. It follows that $\hat{\phi}_n(\eta_\alpha) \leq K_{\mathtt{X}}\ \forall\ \alpha \in \{-1,1\}^d$. Now, using Lemma 2.4.3 we have $\mathtt{X} \subset Conv\left(\eta_\alpha : \alpha \in \{-1,1\}^d\right)$ and the convexity of $\hat{\phi}_n$ implies that $\hat{\phi}_n(x) \leq K_{\mathtt{X}}$ for any $x \in \mathtt{X}$. $\square$

2.4.4 Proof of Lemma 2.3.4

Assume that $\mathtt{X} = [a,b]$ is a rectangle with vertices $\{z_\alpha : \alpha \in \{-1,1\}^d\}$. The function $\psi(x) = \inf_{\eta\in\overline{\mathrm{Ext}(\mathfrak{X})}}\{|x - \eta|\}$ is continuous on $\mathbb{R}^d$ so there is

$x_* \in \partial\mathtt{X}$ such that $\psi(x_*) = \inf_{x\in\partial\mathtt{X}}\{\psi(x)\}$. Observe that $\psi(x_*) > 0$ because $x_* \in \partial\mathtt{X} \subset \mathfrak{X}^\circ$. By Lemma 2.4.3, there is a $r < \frac{1}{2}\psi(x_*)$ for which there exists $\rho < \frac{1}{4}r$ such that whenever $\eta_\alpha \in A_\alpha := B\left(z_\alpha + \frac{3}{4}r\left(\frac{z_\alpha - z_{-\alpha}}{|z_\alpha - z_{-\alpha}|}\right), \rho\right)$ for any $\alpha \in \{-1,1\}^d$ and

$$\begin{aligned} K_z &= Conv\left(z_\alpha + \frac{1}{2}r\left(\frac{z_\alpha - z_{-\alpha}}{|z_\alpha - z_{-\alpha}|}\right) : \alpha \in \{-1,1\}^d\right) \\ K_\eta &= Conv\left(\eta_\alpha : \alpha \in \{-1,1\}^d\right) \end{aligned}$$

we have

$$\mathtt{X} \subset K_z \subset K_\eta^\circ \subset K_\eta \subset \mathfrak{X}^\circ. \tag{2.26}$$

Let $M_0 > \frac{\sigma}{\sqrt{\min\{\nu(A_\alpha):\alpha\in\{-1,1\}^d\}}}$ and $M_1 \in \mathbb{R}$ be such that

$$\mathbf{P}\left(\inf_{x\in\mathtt{X}}\{\hat{\phi}_n(x)\} \le -M_0 \text{ i.o.}\right) = 0 \quad \text{and} \quad M_1 = \sup_{x\in Conv\left(\bigcup_{\alpha\in\{-1,1\}^d} A_\alpha\right)}\{\phi(x)\}.$$

From Lemmas 2.3.1 and 2.3.2 we can find, with probability one, $n_0 \in \mathbb{N}$ such that $\inf_{x\in\mathtt{X}}\{\hat{\phi}_n(x)\} > -M_0$ and $\inf_{x\in A_\alpha}\{|\hat{\phi}_n(x) - \phi(x)|\} < M_0$ for any $n \ge n_0$. Define

$$\begin{aligned} M &= M_1 + M_0 \\ K_{\mathtt{X}} &= \frac{4|b-a|}{r\min_{1\le j\le d}\{b^j - a^j\}}M \end{aligned}$$

and take any $n \ge n_0$. Then, for any $\alpha \in \{-1,1\}^d$ we can find $\eta_\alpha \in A_\alpha$ such that $|\hat{\phi}_n(\eta_\alpha) - \phi(\eta_\alpha)| < M_0$. Then, (2.26) implies that $\hat{\phi}_n(x) \le M\ \forall x \in \mathtt{X}$. Take then $x \in \mathtt{X}$ and $\xi \in \partial\hat{\phi}_n(x)$. A connectedness argument, like the one used in the proof of Lemma 2.3.2, implies that there is $t_* > 0$ such that $x + t_*\xi \in \partial K_\eta$. But then we must have $t_* > \frac{r\min_{1\le j\le d}\{b^j - a^j\}}{2|\xi||b-a|}$ as a consequence of (2.26), since the smallest distance between ∂K_z and $\partial\mathtt{X}$ is $\frac{r\min_{1\le j\le d}\{b^j-a^j\}}{2|b-a|}$ and $\partial K_\eta \subset \text{Ext}(K_z)$. This can be seen by taking a look at Figure 2.4, which shows the situation in the two dimensional case. Thus, using the definition

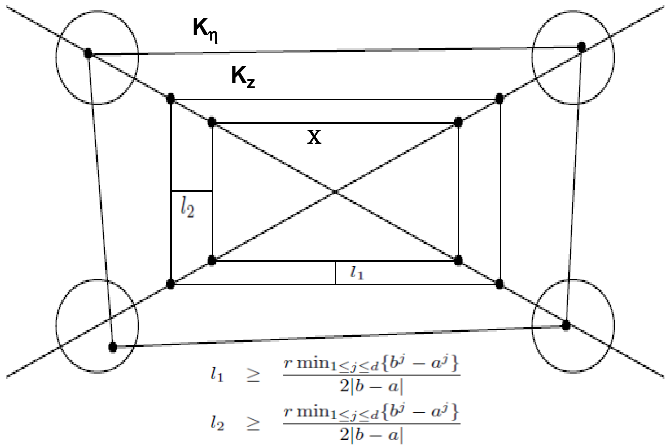

Figure 2.4: The smallest distance between ∂K_z and $\partial \mathtt{X}$ is at least $\frac{r \min_{1\leq j\leq d}\{b^j - a^j\}}{2|b-a|}$.

of subgradients,

$$\frac{r \min_{1\leq j\leq d}\{b^j - a^j\}}{2|\xi||b-a|}\langle \xi, \xi\rangle \leq \langle \xi, t_*\xi\rangle \leq \hat{\phi}_n(x+t_*\xi) - \hat{\phi}_n(x) \leq 2M$$

which in turn implies $|\xi| \leq K_{\mathtt{X}}$. We have therefore shown that, with probability one, we can find $n_0 \in \mathbb{N}$ such that $|\xi| \leq K_{\mathtt{X}}$ $\forall$ $\xi \in \partial\hat{\phi}_n(x)$, $\forall$ $x \in \mathtt{X}$, $\forall$ $n \geq n_0$. This completes the proof. $\square$

2.4.5 Proof of Lemma 2.3.5

This Lemma is a direct consequence of Theorem 6 in Bronšteĭn (1976) (see also Corollary 2.7.10 in page 164 of van der Vaart and Wellner (1996)). Nevertheless, to make this thesis a bit more self-contained, we now present a proof based on elementary computations.

The result is obvious for conditions {A1-A3} and {A5-A7} when $\sigma^2 = 0$. So we assume that $\sigma^2 > 0$ for {A1-A3} and {A5-A7}. Let $\epsilon > 0$ and

$M = \sup_{x \in \mathtt{X}}\{|x|\}$. Choose $\delta > 0$ satisfying

$$\frac{\epsilon}{\frac{2(2M+K\sqrt{d}+1)}{n}\sum_{j=1}^{n}|Y_j - \phi(X_j)|} < \delta < \frac{\epsilon}{\frac{(2M+K\sqrt{d}+1)}{n}\sum_{j=1}^{n}|Y_j - \phi(X_j)|} \tag{2.27}$$

for n large. Notice that δ is well-defined and the quantity on the left is positive, finite and bounded away from 0 as $\underline{\lim}\,\frac{1}{n}\sum_{j=1}^{n}|Y_j - \phi(X_j)| > 0$ a.s. under any set of regularity conditions (for {A2-A4}, conditions A4-(i) and A4-(iii) imply that we can apply the version of the strong law of large number for uncorrelated random variables, as it appears in Chung (2001), page 108, Theorem 5.1.2 to the sequence $(|\epsilon_j|)_{j=1}^{\infty}$; for {A1-A3} and {A5-A7} this is immediate as $\sigma^2 > 0$). The definition of the class $\mathcal{D}_{K,\mathtt{X}}$ implies that all its members are Lipschitz functions with Lipschitz constant bounded by $K\sqrt{d}$, a consequence of Rockafellar (1970), Theorem 24.7, page 237. Hence, (2.27) implies that

$$\sup_{\substack{|x-y|<\delta \\ x,y\in\mathtt{X},\psi\in\mathcal{D}_{K,\mathtt{X}}}}\{|\psi(x)-\psi(y)|\} \leq \frac{\epsilon}{\frac{1}{n}\sum_{j=1}^{n}|Y_j - \phi(X_j)|}.$$

Now, define $N_n \in \mathbb{N}$ by $N_n = \left\lceil\frac{\mathrm{diam}(\mathtt{X})}{\delta}\right\rceil \vee \left\lceil\frac{2K\sqrt{d}}{\delta}\right\rceil$, where $\lceil\cdot\rceil$ denotes the ceiling function. Observe that (2.27) implies

$$N_n - 1 \leq \left(\mathrm{diam}(\mathtt{X}) \vee 2K\sqrt{d}\right)\frac{2(2M+K\sqrt{d}+1)}{\epsilon}\left(\frac{1}{n}\sum_{j=1}^{n}|Y_j - \phi(X_j)|\right). \tag{2.28}$$

Then, we can divide the rectangles $\mathtt{X}$ and $[-K,K]^d$ in N_n^d subrectangles, all of which have diameters less than δ. In other words, we can write

$$\begin{aligned} [-K,K]^d &= \bigcup_{1\leq j\leq N_n^d} R_j \\ \mathtt{X} &= \bigcup_{1\leq j\leq N_n^d} V_j \end{aligned}$$

with $\mathrm{diam}(R_j) < \delta$ and $\mathrm{diam}(V_j) < \delta \ \forall\ j = 1, \dots N_n^d$. In the same way, we can divide the interval $[-K, K]$ in N_n subintervals $\mathcal{I}_1, \dots, \mathcal{I}_{N_n}$ each having length less than δ. For each $j = 1, \dots, N_n^d$, let ξ_j and x_j be the centroids of R_j and V_j respectively and for $j = 1, \dots, N_n$ let η_j be the midpoint of $\mathcal{I}_j$. Consider the class of functions $\mathcal{H}_{n,\epsilon}$ defined by

$$\mathcal{H}_{n,\epsilon} = \left\{ \max_{(s,t,j)\in\mathcal{S}} \{\langle \xi_s, \cdot - x_t\rangle + \eta_j\} : \mathcal{S} \subset \{1, \dots, N_n^d\}^2 \times \{1, \dots, N_n\} \right\}.$$

Observe that the number of elements in the class $\mathcal{H}_{n,\epsilon}$ is bounded from above by $2^{N_n^{2d+1}}$. Now, take any $\psi \in \mathcal{D}_{K,\mathtt{X}}$. Pick any $\Xi_j \in \partial\psi(X_j)$. Then, for any j such that $X_j \in \mathtt{X}$, there are $s_j, t_j \in \{1, \dots, N_n^d\}$ and $\tau_j \in \{1, \dots, N_n\}$ such that $|\Xi_j - \xi_{s_j}|$, $|X_j - x_{t_j}|$ and $|\psi(x_{t_j}) - \eta_{\tau_j}|$ are all less than δ. We then have that

$$\begin{aligned} &\sup_{x\in\mathtt{X}} \left\{ \left| \langle \xi_{s_j}, x - x_{t_j}\rangle + \eta_{\tau_j} - (\langle \Xi_j, x - X_j\rangle + \psi(X_j)) \right| \right\} \\ \leq\ & 2M|\xi_{s_j} - \Xi_j| + K\sqrt{d}|x_{t_j} - X_j| + \delta < (2M + K\sqrt{d} + 1)\delta \end{aligned} \tag{2.29}$$

by an application of the Cauchy-Schwarz inequality. But then, (2.27) implies that if we define the functions $\tilde{\psi}$ and g as

$$\begin{aligned} \tilde{\psi}(x) &= \max_{X_j\in\mathtt{X}} \{\langle \Xi_j, x - X_j\rangle + \psi(X_j)\}, \\ g(x) &= \max_{X_j\in\mathtt{X}} \{\langle \xi_{s_j}, x - x_{t_j}\rangle + \eta_{\tau_j}\} \end{aligned}$$

then we have

$$\tilde{\psi}(X_j) = \psi(X_j) \ \text{ for } j \text{ such that } X_j \in \mathtt{X}, \tag{2.30}$$

$$\|g - \tilde{\psi}\|_{\mathtt{X}} \leq \frac{\epsilon}{\frac{1}{n}\sum_{j=1}^n |Y_j - \phi(X_j)|} \ \text{(from (2.29))}, \tag{2.31}$$

$$g \in \mathcal{H}_{n,\epsilon}. \tag{2.32}$$

Note that (2.30) follows from the definition of subgradients. All these facts put together give that for any $f(x, y) = \psi(x)(y - \phi(x)) \in \mathcal{G}_{K,\mathtt{X}}$, $\psi \in \mathcal{D}_{K,\mathtt{X}}$

there is $g \in \mathcal{H}_{n,\epsilon}$ such that

$$\int_{\mathtt{X}} |f(x,y) - g(x)(y-\phi(x))|\mu_n(dx,dy) < \epsilon$$

and hence

$$N(\epsilon, \mathcal{G}_{K,\mathtt{X}}, \mathbb{L}_1(\mathtt{X}\times\mathbb{R},\mu_n)) \le \#\mathcal{H}_{n,\epsilon} \le 2^{N_n^{2d+1}}.$$

But then, the strong law of large numbers and (2.28) give that $\overline{\lim}\, N_n < \infty$ a.s. Furthermore, by replacing ϵ with $\frac{\epsilon}{n}\sum_{j=1}^n |Y_j - \phi(X_j)|$ in the entire construction just made, we can see that the covering numbers $N\left(\frac{\epsilon}{n}\sum_{j=1}^n |Y_j-\phi(X_j)|, \mathcal{G}_{K,\mathtt{X}}, \mathbb{L}_1(\mathtt{X}\times\mathbb{R},\mu_n)\right)$ depend neither on the Y's nor on ϕ. Taking $B_\epsilon = \left(\mathrm{diam}(\mathtt{X}) \vee K\sqrt{d}\right)\frac{2(2M+K\sqrt{d}+1)}{\epsilon} + 1$ and $A_\epsilon = 2^{B_\epsilon^{2d+1}}$ it is seen that the second part of the result holds. □

2.4.6 Proof of Lemma 2.3.6

Note that for every m, we have

$$\frac{1}{n_k}\sum_{1\le j\le n_k} \mathbf{E}\left(\epsilon_j^2\right) \le \frac{1}{n_k}\sum_{\substack{X_j\in\mathtt{X}_m\\ 1\le j\le n_k}} \mathbf{E}\left(\epsilon_j^2\right) + \frac{N_{n_k}(\mathfrak{X}\setminus\mathtt{X}_m)}{n_k}\sup_{j\in\mathbb{N}}\{\mathbf{E}\left(\epsilon_j^2\right)\}.$$

Taking limit inferior on both sides as $k\to\infty$, we get

$$\sigma^2 \le \varliminf_{k\to\infty}\frac{1}{n_k}\sum_{\substack{X_j\in\mathtt{X}_m\\ 1\le j\le n_k}} \mathbf{E}\left(\epsilon_j^2\right) + \nu(\mathfrak{X}\setminus\mathtt{X}_m)\sup_{j\in\mathbb{N}}\{\mathbf{E}\left(\epsilon_j^2\right)\}.$$

Now taking the limit as $m\to\infty$ we get the result because the opposite inequality is trivial. □

2.4.7 Proof of Lemma 2.3.7

We may assume that $\mathtt{X}$ is a compact rectangle. Here we need to make a distinction between the design schemes. In the case of the stochastic design, the proof is an immediate consequence of Lemma 2.3.5 and Theorem 2.4.3, page 123 of van der Vaart and Wellner (1996). Thus, we focus on the fixed design scenario.

For notational convenience, we write $M = \sup_{j\in\mathbb{N}}\{\mathbf{E}\left(\epsilon_j^2\right)\}$ and $\sum_{X_j\in\mathtt{X}}$ instead of the more cumbersome $\sum_{1\le j\le n: X_j\in\mathtt{X}}$. Letting $\epsilon_j = Y_j - \phi(X_j)$ (and using the same notation as in the proof of Lemma 2.3.7) first observe that the random quantity

$$\sup_{\psi\in\mathcal{D}_{K,\mathtt{X}}}\left\{\left|\frac{1}{n}\sum_{\{X_j\in\mathtt{X}\}}\psi(X_j)\epsilon_j\right|\right\} = \sup_{m\in\mathbb{N}}\left\{\sup_{g\in\mathcal{H}_{n,\frac{1}{m}}}\left\{\left|\frac{1}{n}\sum_{\{X_j\in\mathtt{X}\}}g(X_j)\epsilon_j\right|\right\}\right\}.$$

by (2.30), (2.31) and (2.32) and is thus measurable.

All of the following arguments are valid for both, {A1-A3} and {A2-A4}. Lyapunov's inequality (which states that for any random variable X and $1\le p\le q\le\infty$ we have $\|X\|_p \le \|X\|_q$) and the strong law of large numbers imply

$$\overline{\lim_{m\to\infty}}\frac{1}{m}\sum_{1\le j\le m}|\epsilon_j| = \overline{\lim_{m\to\infty}}\frac{1}{m}\sum_{1\le j\le m}\mathbf{E}\left(|\epsilon_j|\right)\le\sqrt{M}\text{ a.s.}\tag{2.33}$$

Let $\eta > 0$. From Lemma 2.3.5 we know that the covering numbers $a_n := N\left(\frac{\eta}{n}\sum_{j=1}^n|Y_j-\phi(X_j)|, \mathcal{G}_{K,\mathtt{X}}, \mathbb{L}_1(\mathtt{X}\times\mathbb{R},\mu_n)\right)$ are not random and uniformly bounded by a constant A_η. Therefore, for any $n\in\mathbb{N}$ we can find a class $\mathcal{A}_n\subset\mathcal{D}_{K,\mathtt{X}}$ with exactly a_n elements such that $\{\psi(x)(y-\phi(x))\}_{\psi\in\mathcal{A}_n}$ forms an $\left(\frac{\eta}{n}\sum_{j=1}^n|Y_j-\phi(X_j)|\right)$-net for $\mathcal{G}_{K,\mathtt{X}}$ with respect to $\mathbb{L}_1(\mathtt{X}\times\mathbb{R},\mu_n)$. It follows that

$$\sup_{\psi\in\mathcal{D}_{K,\mathtt{X}}}\left\{\left|\frac{1}{n}\sum_{X_j\in\mathtt{X}}\psi(X_j)\epsilon_j\right|\right\}\le\frac{\eta}{n}\sum_{1\le j\le n}|\epsilon_j| + \sup_{\psi\in\mathcal{A}_n}\left\{\left|\frac{1}{n}\sum_{X_j\in\mathtt{X}}\psi(X_j)\epsilon_j\right|\right\}.\tag{2.34}$$

With (2.34) in mind, we make the following definitions

$$
\begin{aligned}
B_n &= \sup_{\psi\in\mathcal{A}_n}\left\{\left|\frac{1}{n}\sum_{X_j\in\mathtt{X}}\psi(X_j)\epsilon_j\right|\right\},\\
C_n &= \sup_{\psi\in\mathcal{A}_n}\left\{\left|\frac{1}{n}\sum_{1\le j\le\lfloor\sqrt{n}\rfloor^2:\ X_j\in\mathtt{X}}\psi(X_j)\epsilon_j\right|\right\},\\
D_n &= \sup_{\substack{\psi\in\mathcal{A}_k\\ n^2\le k<(n+1)^2}}\left\{\left|\frac{1}{k}\sum_{n^2<j\le k:\ X_j\in\mathtt{X}}\psi(X_j)\epsilon_j\right|\right\},
\end{aligned}
$$

where $\lfloor\cdot\rfloor$ denotes the floor function. Now, pick $\delta>0$ and observe that

$$
\begin{aligned}
\mathbf{P}\left(B_n>\delta\right) &= \mathbf{P}\left(\bigcup_{\psi\in\mathcal{A}_n}\left[\left|\sum_{X_j\in\mathtt{X}}\psi(X_j)\epsilon_j\right|>n\delta\right]\right)\\
&\le \sum_{\psi\in\mathcal{A}_n}\frac{1}{n^2\delta^2}M\sum_{X_j\in\mathtt{X}}\psi(X_j)^2 \;\le\; \frac{K^2MA_\eta}{n\delta^2}.
\end{aligned}
$$

The Borel-Cantelli Lemma then implies that $\mathbf{P}\left(B_{n^2}>\delta \text{ i.o.}\right)=0$. Letting $\delta\to 0$ through a decreasing sequence gives

$$
B_{n^2}\xrightarrow{a.s.}0. \tag{2.35}
$$

On the other hand, the definition of C_n implies that

$$
C_n\le\frac{\lfloor\sqrt{n}\rfloor^2}{n}B_{\lfloor\sqrt{n}\rfloor^2}+\frac{\eta}{n}\sum_{1\le j\le\lfloor\sqrt{n}\rfloor^2}|\epsilon_j| \tag{2.36}
$$

which together with (2.35) and (2.33) gives

$$
\overline{\lim}\, C_n\le\eta\sqrt{M}\quad\text{almost surely.} \tag{2.37}
$$

Note that (2.36) is a consequence of the fact that for any $\psi\in\mathcal{A}_n$, there exists $g\in\mathcal{A}_{\lfloor\sqrt{n}\rfloor^2}$ such that if $\mathcal{J}_n=\{1\le j\le\lfloor\sqrt{n}\rfloor^2 : X_j\in\mathtt{X}\}$, then

$$
\begin{aligned}
\left|\frac{1}{n}\sum_{j\in\mathcal{J}_n}\psi(X_j)\epsilon_j\right| &\le \left|\frac{1}{n}\sum_{j\in\mathcal{J}_n}(\psi(X_j)-g(X_j))\epsilon_j\right|+\left|\frac{1}{n}\sum_{j\in\mathcal{J}_n}g(X_j)\epsilon_j\right|\\
&\le \left(\frac{\lfloor\sqrt{n}\rfloor^2}{n}\right)\frac{\eta}{\lfloor\sqrt{n}\rfloor^2}\sum_{1\le j\le\lfloor\sqrt{n}\rfloor^2}|\epsilon_j|+\frac{\lfloor\sqrt{n}\rfloor^2}{n}B_{\lfloor\sqrt{n}\rfloor^2}.
\end{aligned}
$$

Now, a similar argument to the one used in (2.35) gives

$$
\begin{aligned}
\mathbf{P}\left(D_n > \delta\right) &= \mathbf{P}\left(\bigcup_{\substack{\psi\in\mathcal{A}_k\\ n^2\le k<(n+1)^2}}\left[\left|\sum_{n^2<j\le k: X_j\in\mathtt{X}}\psi(X_j)\epsilon_j\right| > k\delta\right]\right)\\
&\le \sum_{\substack{\psi\in\mathcal{A}_k\\ n^2\le k<(n+1)^2}} \mathbf{P}\left(\left|\sum_{n^2<j\le k: X_j\in\mathtt{X}}\psi(X_j)\epsilon_j\right| > k\delta\right)\\
&\le \sum_{\substack{\psi\in\mathcal{A}_k\\ n^2\le k<(n+1)^2}} \frac{K^2M(k-n^2)}{k^2\delta^2} \le \frac{K^2MA_\eta(2n+1)^2}{n^4\delta^2} \qquad (2.38)
\end{aligned}
$$

Again, one can use (2.38) and the Borel-Cantelli Lemma to prove that $\mathbf{P}\left(D_n > \delta \text{ i.o.}\right) = 0$ and then let $\delta \to 0$ through a decreasing sequence to obtain

$$D_n \xrightarrow{a.s.} 0. \qquad (2.39)$$

Finally, one sees that

$$\sup_{\psi\in\mathcal{A}_n}\left\{\left|\frac{1}{n}\sum_{X_j\in\mathtt{X}}\psi(X_j)(Y_j-\phi(X_j))\right|\right\} = B_n \le C_n + D_{\lfloor\sqrt{n}\rfloor},$$

which combined with (2.37) and (2.39) gives

$$\overline{\lim}\, B_n \le \eta\sqrt{M} \quad \text{almost surely.}$$

Taking (2.34) into account we get

$$\overline{\lim_{n\to\infty}}\sup_{\psi\in\mathcal{D}_{K,\mathtt{X}}}\left\{\left|\frac{1}{n}\sum_{1\le j\le n: X_j\in\mathtt{X}}\psi(X_j)(Y_j-\phi(X_j))\right|\right\} \le 2\eta\sqrt{M} \quad \text{almost surely.}$$

Letting $\eta \to 0$ we get the desired result. $\square$

2.4.8 Proof of Lemma 2.3.8

We can assume, without loss of generality, that $\mathtt{X}$ is a finite union of compact rectangles. Consider a sequence $(\mathtt{X}_m)_{m=1}^{\infty}$ satisfying the following properties:

(a) $\mathtt{X} \subset \mathtt{X}_m \subset \mathfrak{X}^\circ \ \forall\ m \in \mathbb{N}$.

(b) $\nu(\mathtt{X}_m) > 1 - \frac{1}{m} \ \forall\ m \in \mathbb{N}$.

(c) $\mathtt{X}_m \subset \mathtt{X}_{m+1} \ \forall\ m \in \mathbb{N}$.

(d) Every $\mathtt{X}_m$ can be expressed as a finite union of compact rectangles with positive Lebesgue measure.

The existence of such a sequence follows from the inner regularity of Borel probability measures on $\mathbb{R}^d$ and from the fact that since $\mathfrak{X}^\circ$ is open, for any compact set $F \subset \mathfrak{X}^\circ$ we can find a finite cover composed by compact rectangles with positive Lebesgue measure and completely contained in $\mathfrak{X}^\circ$. Also, from Lemmas 2.3.2, 2.3.3 and 2.3.4 and the fact that $\mathfrak{X} \subset dom(\phi)$, for any $m \in \mathbb{N}$ we can find $K_m > 0$ such that

$$\|\phi\|_{\mathtt{X}_m} \leq K_m \quad \text{and} \quad \mathbf{P}\left(\|\hat{\phi}_n\|_{\mathtt{X}_m} > K_m \ \text{ i.o.}\right) = 0; \tag{2.40}$$

$$\sup_{\substack{x \in \mathtt{X}_m \\ \xi \in \partial\phi(x)}} \{|\xi|\} \leq K_m \quad \text{and} \quad \mathbf{P}^*\left(\sup_{\substack{x \in \mathtt{X}_m \\ \xi \in \partial\hat{\phi}_n(x)}} \{|\xi|\} > K_m \ \text{ i.o.}\right) = 0. \tag{2.41}$$

Fix $\eta > 0$ and consider the sets

$$\begin{aligned} A &= \left[\inf_{x \in \mathtt{X}} \{\phi(x) - \hat{\phi}_n(x)\} \geq \eta \ \text{ i.o.}\right] \\ B &= \left[\|\hat{\phi}_n\|_{\mathtt{X}_m} \leq K_m \ \text{ a.a.}\right] \\ C &= \left[\sup_{\substack{x \in \mathtt{X}_m \\ \xi \in \partial\hat{\phi}_n(x)}} \{|\xi|\} \leq K_m \ \text{ a.a.}\right]. \end{aligned}$$

Suppose now that $A \cap B \cap C$ is known to be true. Then, there is a subsequence $(n_k)_{k=1}^{\infty}$ such that $\inf_{x \in \mathtt{X}}\{\phi(x) - \hat{\phi}_{n_k}(x)\} \geq \eta \ \forall\ k \in \mathbb{N}$ and $\frac{1}{n_k}\sum_{j=1}^{n_k} \mathbf{E}\left(\epsilon_j^2\right) \to \sigma^2$.

Taking (2.40) and (2.41) into account, we have that for k large enough the inequality

$$\frac{1}{n_k}\sum_{j=1}^{n_k}(Y_j-\hat{\phi}_{n_k}(X_j))^2 \geq \frac{1}{n_k}\sum_{X_j\in \mathtt{X}_m}(Y_j-\phi(X_j))^2$$

$$+\frac{2}{n_k}\sum_{X_j\in \mathtt{X}_m}(Y_j-\phi(X_j))(\phi(X_j)-\hat{\phi}_{n_k}(X_j))+\frac{1}{n_k}\sum_{X_j\in \mathtt{X}_m}(\phi(X_j)-\hat{\phi}_{n_k}(X_j))^2$$

implies

$$\frac{1}{n_k}\sum_{j=1}^{n_k}(Y_j-\hat{\phi}_{n_k}(X_j))^2 \geq \frac{1}{n_k}\sum_{X_j\in \mathtt{X}_m}(Y_j-\phi(X_j))^2 +$$

$$\frac{N_{n_k}(\mathtt{X})}{n_k}\eta^2 - 4\sup_{\psi\in\mathcal{D}_{K_m,\mathtt{X}_m}}\left\{\left|\frac{1}{n_k}\sum_{\{1\leq j\leq n_k: X_j\in\mathtt{X}_m\}}\psi(X_j)(Y_j-\phi(X_j))\right|\right\}.$$

Thus, from Lemma 2.3.7 we can conclude that

$$\varliminf_{k\to\infty}\frac{1}{n_k}\sum_{1\leq j\leq n_k}(Y_j-\hat{\phi}_{n_k}(X_j))^2 \geq \nu(\mathtt{X}_m)\sigma^2+\nu(\mathtt{X})\eta^2 \text{ if } \{\text{A1-A3}\} \text{ hold.}$$

Under {A2-A4} and {A5-A7} the left-hand side of the last display is bounded from below by

$$\varliminf_{k\to\infty}\frac{1}{n_k}\sum_{X_j\in\mathtt{X}_m}(Y_j-\phi(X_j))^2+\nu(\mathtt{X})\eta^2$$

and

$$\int_{\mathtt{X}_m}(y-\phi(x))^2\mu(dx,dy)+\nu(\mathtt{X})\eta^2,$$

respectively.

Finally, using (a)-(d), the strong law of large numbers (for {A2-A4} we can apply a version of the strong law of large numbers for independent random variables thanks to condition A4-(ii); see Williams (1991), Lemma 12.8, page 118 or Folland (1999), Theorem 10.12, page 322) and Lemma 2.3.6 we can let $m\to\infty$ to see that, under any of {A1-A3}, {A2-A4} or {A5-A7},

$$\varliminf_{k\to\infty}\frac{1}{n_k}\sum_{1\leq j\leq n_k}(Y_j-\hat{\phi}_{n_k}(X_j))^2 \geq \sigma^2+\nu(\mathtt{X})\eta^2$$

which is impossible because $\hat{\phi}_{n_k}$ is the least squares estimator.

Therefore $\mathbf{P}^*(A \cap B \cap C) = 0$ and, since $\mathbf{P}_*(B \cap C) = 1$,

$$\mathbf{P}(A) = \mathbf{P}\left(\inf_{x \in \mathtt{X}} \{\phi(x) - \hat{\phi}_n(x)\} \geq \eta \text{ i.o.}\right) = 0.$$

This finishes the proof of (i). The second assertion follows from similar arguments. □

2.4.9 Proof of Lemma 2.3.9

We can assume, without loss of generality, that $\mathtt{X}$ is a finite union of compact rectangles. Pick $K_{\mathtt{X}}$ such that

$$\sup_{\substack{x \in \mathtt{X} \\ \xi \in \partial\phi(x)}} \{|\xi|\} \leq K_{\mathtt{X}} \text{ and } \mathbf{P}^*\left(\sup_{\substack{x \in \mathtt{X} \\ \xi \in \partial\hat{\phi}_n(x)}} \{|\xi|\} > K_{\mathtt{X}} \text{ i.o.}\right) = 0.$$

Let $\eta > 0$ and $\delta = \frac{\eta}{3K_{\mathtt{X}}}$. We can then divide $\mathtt{X}$ in M subrectangles $\{\mathcal{C}_1, \ldots, \mathcal{C}_M\}$ all having diameter less than δ. Define the events

$$\begin{aligned} A &= \left[\bigcap_{1 \leq k \leq M} \inf_{x \in \mathcal{C}_k} \{\hat{\phi}_n(x) - \phi(x)\} < \frac{\eta}{3} \text{ a.a.}\right] \\ B &= \left[\sup_{\substack{x \in \mathtt{X} \\ \xi \in \partial\hat{\phi}_n(x)}} \{|\xi|\} \leq K_{\mathtt{X}} \text{ a.a.}\right]. \end{aligned}$$

We will show that $A \cap B \subset \left[\sup_{x \in \mathtt{X}} \{\hat{\phi}_n(x) - \phi(x)\} \leq \eta \text{ a.a.}\right]$. Suppose $A \cap B$ is true. Then, there is $N \in \mathbb{N}$ such that for any $n \geq N$ we can find $\Xi_{n,k} \in \mathcal{C}_k$ such that $\hat{\phi}_n(\Xi_{n,k}) - \phi(\Xi_{n,k}) < \frac{\eta}{3}$. Moreover, we can make N large enough such that for any $n \geq N$, $K_{\mathtt{X}}$ is an upper bound for all the subgradients of $\hat{\phi}_n$ on $\mathtt{X}$. Then, for any $\xi \in \mathcal{C}_k$ we obtain from the Lipschitz property,

$$\begin{aligned} \hat{\phi}_n(\xi) - \phi(\xi) &= (\hat{\phi}_n(\Xi_{n,k}) - \phi(\Xi_{n,k})) + (\phi(\Xi_{n,k}) - \phi(\xi)) + (\hat{\phi}_n(\xi) - \hat{\phi}_n(\Xi_{n,k})) \\ &\leq \frac{\eta}{3} + K_{\mathtt{X}}\delta + K_{\mathtt{X}}\delta \leq \eta. \end{aligned}$$

Therefore,

$$\sup_{x\in\mathcal{C}_k}\{\hat{\phi}_n(x)-\phi(x)\}\le\eta \quad \forall\, 1\le k\le M\ \forall\, n\ge N$$

which implies

$$\sup_{x\in\mathtt{X}}\{\hat{\phi}_n(x)-\phi(x)\}\le\eta \quad \forall\, n\ge N.$$

Considering Lemmas 2.3.8-(ii) and 2.3.4; $A\cap B\subset\left[\sup_{x\in\mathtt{X}}\{\hat{\phi}_n(x)-\phi(x)\}\le\eta\ \text{ a.a.}\right]$ and $\mathbf{P}_*\left(A\cap B\right)=1$ we obtain (ii). The first assertion follows from similar arguments and (iii) is a direct consequence of (i) and (ii). □

2.4.10 Proof of Lemma 2.3.10

Throughout this proof we will denote by $\mathbf{B}$ the unit ball (w.r.t. the euclidian norm) in $\mathbb{R}^d$. From Theorem 25.5, page 246 on Rockafellar (1970) we know that f is continuously differentiable on $\mathcal{C}$. Let

$$h_*=\inf_{\xi\in\mathtt{X},\eta\in\mathbb{R}^d\setminus\mathcal{C}}\{|\xi-\eta|\}>0.$$

Pick $\epsilon>0$. We will first show that there is $n_\epsilon\in\mathbb{N}$ such that

$$\langle\xi,\eta\rangle\le\langle\nabla f(x),\eta\rangle+\epsilon,\quad \forall\,\xi\in\partial f_n(x),\ \forall\,x\in\mathtt{X},\ \forall\,\eta\in\mathbf{B},\ \forall\,n\ge n_\epsilon. \tag{2.42}$$

Suppose that such an n_ϵ does not exist. Then, there is an increasing sequence $(m_n)_{n=1}^\infty$ such that for any $n\in\mathbb{N}$ we can find $x_{m_n}\in\mathtt{X}$, $\xi_{m_n}\in\partial f_{m_n}(x_{m_n})$, $\eta_{m_n}\in\mathbf{B}$ satisfying $\langle\xi_{m_n},\eta_{m_n}\rangle>\langle\nabla f(x_{m_n}),\eta_{m_n}\rangle+\epsilon$. But $\mathtt{X}$ and $\mathbf{B}$ are both compact, so there are $x_*\in\mathtt{X}$, $\eta_*\in\mathbf{B}$ and a subsequence $(k_n)_{n=1}^\infty$ of $(m_n)_{n=1}^\infty$ such that $x_{k_n}\to x_*$ and $\eta_{k_n}\to\eta_*$. Then, for any $0<h<h_*$ we have

$$\frac{f_{k_n}(x_{k_n}+h\eta_{k_n})-f_{k_n}(x_{k_n})}{h}\ge\langle\xi_{k_n},\eta_{k_n}\rangle>\langle\nabla f(x_{m_n}),\eta_{k_n}\rangle+\epsilon\ \ \forall\,n\in\mathbb{N},$$

and therefore

$$\varliminf_{n\to\infty}\lim_{h\downarrow 0}\frac{f_{k_n}(x_{k_n}+h\eta_{k_n})-f_{k_n}(x_{k_n})}{h}\ge\langle\nabla f(x_*),\eta_*\rangle+\epsilon.$$

But this is impossible in view of Theorem 24.5, page 233 on Rockafellar (1970). It follows that we can choose some $n_\epsilon \in \mathbb{N}$ with the property described in (2.42). By noting that $-\mathbf{B} = \mathbf{B}$, we can conclude from (2.42) that

$$|\langle \xi, \eta \rangle - \langle \nabla f(x), \eta \rangle| \leq \epsilon \quad \forall\ \xi \in \partial f_n(x),\ \forall\ x \in \mathtt{X},\ \forall\ \eta\ \in \mathbf{B},\ \forall\ n \geq n_\epsilon.$$

By taking $\eta_\xi = \frac{\xi - \nabla f(x)}{|\xi - \nabla f(x)|}$ when $\xi \neq \nabla f(x)$ we get

$$\sup_{\substack{x \in \mathtt{X} \\ \xi \in \partial f_n(x)}} \{|\xi - \nabla f(x)|\} \leq \epsilon \quad \forall\ n \geq n_\epsilon.$$

Since $\epsilon > 0$ was arbitrarily chosen, this completes the proof. $\square$

Chapter 3

Additional topics regarding convex regression

3.1 The one-dimensional case

In this section we elaborate a bit more on what is known about the convex regression problem with a unidimensional predictor. For simplicity, we assume no ties among the X's. As stated before, a function $f : \mathbb{R} \to \overline{\mathbb{R}}$ is convex if it is absolutely continuous in $dom(f)^{\circ}$ with a nondecreasing first derivative. This fact yields the following characterization of the set $\mathcal{K}_{\mathcal{X}}$ defined in Section 2.2:

$$\mathcal{K}_{\mathcal{X}} := \left\{ z \in \mathbb{R}^n : \frac{z^{(j)} - z^{(j-1)}}{X_{(j)} - X_{(j-1)}} \leq \frac{z^{(j+1)} - z^{(j)}}{X_{(j+1)} - X_{(j)}} \ \forall\ j = 2, \ldots, n-1 \right\}, \quad (3.1)$$

where $X_{(1)} \leq \ldots \leq X_{(n)}$ are the *order statistics* and each $z^{(j)}$ represents the component of z that corresponds to the index k for which $X_k = X_{(j)}$. The vector Z_n defined in (2.2) can now be computed by solving the program

$$\begin{array}{cc} \min & \sum_{k=1}^{n} |Y_k - z^k|^2 \\ \text{subject to} & \frac{z^{(j)} - z^{(j-1)}}{X_{(j)} - X_{(j-1)}} \leq \frac{z^{(j+1)} - z^{(j)}}{X_{(j+1)} - X_{(j)}} \ \forall\ j = 2, \ldots, n-1. \end{array} \quad (3.2)$$

Moreover, the value of $\hat{\phi}_n(x)$ for any $x \in Conv\,(X_1, \ldots, X_n) = [X_{(1)}, X_{(n)}]$ can be obtained by *linearly interpolating* the points $(X_{(1)}, Z^{(1)}), \ldots, (X_{(n)}, Z^{(n)})$.

A direct comparison between (3.1), (3.2) and the interpolating procedure against their multidimensional counterparts might be illustrative: as opposed to the simple, straightforward characterization given by (3.1), Lemmas 2.2.1 and 2.2.2 are much more complex and require the solution of linear programs; the positive definite quadratic program (3.2) is computationally a lot simpler than the semidefinite problem (2.8); evaluating $\hat{\phi}_n$ via linear interpolation is evidently more convenient than solving (2.7).

All these simplifications are consequence of the aforementioned characterization of convex functions on the real line. This fact also plays a key role in the derivation of the rate of convergence and asymptotic distribution. The rate of convergence was shown to be $n^{-2/5}$ in Mammen (1991). Groeneboom et al. (2001) showed that for any $x \in \mathfrak{X}^\circ$ on which ϕ is twice continuously differentiable with $\phi''(x) \neq 0$,

$$n^{2/5}\kappa_{x,\phi}(\hat{\phi}_n(x) - \phi(x)) \rightsquigarrow H''(0)$$

where H is a (well-defined) random continuous function majorizing

$$\int_0^t W(s)ds + t^4$$

with W being a standard, two-sided Brownian motion, and

$$\kappa_{x,\phi} = \left(\frac{24}{\mathbf{Var}\,(\epsilon)^2\,\phi''(x)}\right)^{1/5}.$$

3.2 Convex and componentwise monotone regression functions

As explained in Section 2.1, concave regression arises naturally in econometrics. Production and utility functions are often assumed not only concave, but also componentwise nondecreasing. In this section we adapt the method of least squares for such situations. As there is no mathematical reason to restrict ourselves to concave and componentwise nondecreasing functions, we will expand the framework of Chapter 2 to functions that are convex and monotone (nondecreasing or nonincreasing) in a given set of directions. Considering the notation of Section 2.2, we make the following definition:

Definition 3.2.1 (α-monotone function) *Let $\alpha \in \{-1,0,1\}^d$. A function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is said to be α-monotone if $f(x) \leq f(x + r\alpha^j \boldsymbol{e}_j)$ for all $r \geq 0$, $x \in \mathbb{R}^d$ and $j \in \{1, \ldots, d\}$, where $\mathbf{e}_j$ is the unit vector in the j-th direction (i.e., the vector with components $\mathbf{e}_j^k = \delta_{kj}$ with δ_{kj} denoting the Kronecker δ).*

In other words, $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is α-monotone if it is nondecreasing in all components j for which $\alpha^j = 1$ and nonincreasing in those for which $\alpha^j = -1$. Additionally, we write $\mathbb{R}^d_+$ and $\mathbb{R}^d_-$ for the nonnegative and nonpositive orthants, respectively, and for any $\alpha \in \{-1,0,1\}^d$ define

$$\mathbb{R}^d_\alpha := \left\{x \in \mathbb{R}^d : \alpha^j x^j \geq 0 \text{ if } |\alpha^j| = 1 \text{ and } x^j = 0 \text{ if } \alpha^j = 0\right\}; \tag{3.3}$$

$$\widetilde{\mathbb{R}}^d_\alpha := \left\{x \in \mathbb{R}^d : \alpha^j x^j \geq 0 \text{ if } |\alpha^j| = 1\right\}. \tag{3.4}$$

For example, if $\alpha_1 := (-1,0,1)$, $\alpha_2 := (1,1,1)$ and $\alpha_3 := (0,1,0)$, then $\mathbb{R}^3_{\alpha_1} = \mathbb{R}_- \times \{0\} \times \mathbb{R}_+$, $\mathbb{R}^3_{\alpha_2} = \mathbb{R}^3_+$ and $\mathbb{R}^3_{\alpha_3} = \{0\} \times \mathbb{R}_+ \times \{0\}$. Accordingly we would also have $\widetilde{\mathbb{R}}^3_{\alpha_1} = \mathbb{R}_- \times \mathbb{R} \times \mathbb{R}_+$, $\widetilde{\mathbb{R}}^3_{\alpha_2} = \mathbb{R}^3_+$ and $\widetilde{\mathbb{R}}^3_{\alpha_3} = \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}$.

Remark: A function f will be α-monotone if and only if $f(x) \leq f(y)$ whenever $(y - x) \in \mathbb{R}^d_\alpha$. For a proof of this statement, see Lemma A.0.7.

3.2.1 The convex, α-monotone least squares estimator: computation and finite sample properties

For the remainder of this section, we will fix an $\alpha \in \{-1,0,1\}^d$ and take the regression function ϕ in (2.1) to be convex and α-monotone. The developments here are quite similar to those in Chapter 2, so we omit some of the details. Given the observed values $(X_1, Y_1), \ldots, (X_n, Y_n)$, we write $\mathcal{Q}_{\mathcal{X}}^{\alpha}$ for the collection of all vectors $z \in \mathbb{R}^n$ for which there is a convex, α-monotone function ψ satisfying $\psi(X_j) = z^j$ for every $j = 1, \ldots, n$. We now have the following characterizations.

Lemma 3.2.1 *Let $z \in \mathbb{R}^n$. Then, $z \in \mathcal{Q}_{\mathcal{X}}^{\alpha}$ if and only if the following holds for every $j = 1, \ldots, n$:*

$$z^j = \inf\left\{\sum_{k=1}^{n} \theta^k z^k : \sum_{k=1}^{n} \theta^k = 1,\ \vartheta + \sum_{k=1}^{n} \theta^k X_k = X_j,\ \theta \geq 0,\ \theta \in \mathbb{R}^n, \vartheta \in \mathbb{R}^d_{-\alpha}\right\}.$$

Proof: The proof is very similar to that of Lemma 2.2.1. The difference being that we use Lemma A.0.8 and the function

$$h_\alpha(x) = \inf\left\{\sum_{k=1}^{n} \theta^k z^k : \sum_{k=1}^{n} \theta^k = 1,\ \vartheta + \sum_{k=1}^{n} \theta^k X_k = x,\ \theta \geq 0,\ \theta \in \mathbb{R}^n, \vartheta \in \mathbb{R}^d_{-\alpha}\right\}$$

instead of Lemma A.0.6 and the function g. $\square$

The dual characterization analogous to that in Lemma 2.2.2 is given in the following result. Its proof is just an application of the duality theorem of linear programming, so we omit it. Recall the definition of $\mathbb{R}^d_\alpha$ and $\widetilde{\mathbb{R}}^d_\alpha$ given in (3.3) and (3.4).

Lemma 3.2.2 *Let $z \in \mathbb{R}^n$. Then, $z \in \mathcal{Q}_{\mathcal{X}}^{\alpha}$ if and only if for every $j = 1, \ldots, n$ we have*

$$z^j = \sup\left\{\langle \xi, X_j \rangle + \eta : \langle \xi, X_k \rangle + \eta \leq z^k\ \forall\ k = 1, \ldots, n,\ \xi \in \widetilde{\mathbb{R}}^d_\alpha,\ \eta \in \mathbb{R}\right\}.$$

Moreover, $z \in \mathcal{Q}_{\mathcal{X}}^{\alpha}$ if and only if there exist vectors $\xi_1, \ldots, \xi_n \in \widetilde{\mathbb{R}}^d_\alpha$ such that

$$\langle \xi_j, X_k - X_j \rangle \leq z^k - z^j \quad \forall\ k, j \in \{1, \ldots, n\}.$$

Just as in the previous case, we can use both characterizations to show the existence and uniqueness of the vector

$$W_n = \underset{z\in\mathcal{Q}_{\mathcal{X}}^{\alpha}}{\operatorname{argmin}} \left\{ \sum_{k=1}^{n} \left|Y_k - z^k\right|^2 \right\}$$

and then define the nonparametric least squares estimator by

$$\hat{\varphi}_n(x) = \inf\left\{ \sum_{k=1}^{n} \theta^k W_n^k : \sum_{k=1}^{n} \theta^k = 1, \vartheta + \sum_{k=1}^{n} \theta^k X_k = x, \theta \in \mathbb{R}^n, \vartheta \in \mathbb{R}^d_{-\alpha} \right\}.$$

Here, the vector W_n can also be computed by solving the corresponding quadratic program

$$\begin{array}{rc} \min & \sum_{k=1}^{n} |Y_k - z^k|^2 \\ \text{subject to} & \langle \xi_k, X_j - X_k\rangle \leq z^j - z^k \quad \forall\ k, j = 1, \ldots, n \\ & \xi_1, \ldots, \xi_n \in \widetilde{\mathbb{R}}^d_{\alpha}, z \in \mathbb{R}^n. \end{array}$$

which differs from the program (2.8) just because here the ξ_j's have sign restrictions. The estimator enjoys analogous finite dimensional properties to those listed in Lemma 2.2.4. For the sake of completeness, we include them in the following lemma.

Lemma 3.2.3 *Let $\hat{\varphi}_n$ be the convex, α-monotone least squares estimator obtained from the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Then,*

(i) $\sum_{k=1}^{n} (\psi(x_k) - \hat{\varphi}_n(X_k))(Y_k - \hat{\varphi}_n(X_k)) \leq 0$ *for any convex, α-monotone function ψ which is finite on Conv $(X_1, \ldots, X_n)$;*

(ii) $\sum_{k=1}^{n} \hat{\varphi}_n(X_k)(Y_k - \hat{\varphi}_n(X_k)) = 0$*;*

(iii) $\sum_{k=1}^{n} Y_k = \sum_{k=1}^{n} \hat{\varphi}_n(X_k)$*;*

(iv) the set on which $\hat{\varphi}_n < \infty$ is Conv $(X_1, \ldots, X_n) + \mathbb{R}^d_{-\alpha}$;

(v) for any $x \in \mathbb{R}^d$ the map $(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \mapsto \hat{\varphi}_n(x)$ is a Borel-measurable function from $\mathbb{R}^{n(d+1)}$ into $\mathbb{R}$.

3.2.2 The convex, α-monotone least squares estimator: consistency

With similar arguments to those used in Section 2.3 it can be shown that the convex, α-monotone least squares estimator and its subdifferentials are consistent. A careful analysis of their respective proofs shows that Theorems 2.3.1 and 2.3.2 still hold with $\hat{\phi}_n$ replaced by $\hat{\varphi}_n$ when ϕ is convex and α-monotone. These proofs relied essentially on two key facts:

(i) The finite sample properties of $\hat{\phi}_n$ established in Lemma 2.2.4.

(ii) The vector $(\hat{\phi}_n(X_1), \ldots, \hat{\phi}_n(X_n))' \in \mathbb{R}^n$ is the $\mathcal{L}_2$ projection of $(Y_1, \ldots, Y_n)$ on the closed, convex cone $\mathcal{K}_{\mathcal{X}}$ of all evaluations of proper convex functions on $(X_1, \ldots, X_n)$. Also, note that $(\phi(X_1), \ldots, \phi(X_n))' \in \mathcal{K}_{\mathcal{X}}$.

We know from Lemma 3.2.3 that $\hat{\varphi}_n$ has similar finite sample properties to those of its convex counterpart. Note that if ϕ is convex and α-monotone, $(\phi(X_1), \ldots, \phi(X_n))' \in \mathcal{Q}^{\alpha}_{\mathcal{X}}$ and $(\hat{\varphi}_n(X_1), \ldots, \hat{\varphi}_n(X_n))' \in \mathbb{R}^n$ is the $\mathcal{L}_2$ projection of $(Y_1, \ldots, Y_n)$ onto $\mathcal{Q}^{\alpha}_{\mathcal{X}}$. Finally, as the space of convex and α-monotone functions on $\mathcal{X}$ is a subspace of the space of convex functions on $\mathcal{X}$, the entropy bounds derived in Lemma 2.3.5 still hold.

From these considerations and the nature of the arguments used to prove Theorems 2.3.1 and 2.3.2, it follows that the consistency theorems 2.3.1 and 2.3.2 remain true for the convex, α-monotone least squares estimator.

3.3 Behavior under misspecified model

In this section we analyze behavior of the convex least squares estimator on scenarios in which the actual regression function is *not* convex. It turns out that the estimator has asymptotic regularity in the sense that even in such cases there is a well-defined convex function to which it converges almost surely. The arguments presented here are based on those in Lim and Glynn (2012).

Consider the following situation: we are given an i.i.d. sequence $(X_n, Y_n)_{n=1}^{\infty}$ from some Borel probability measure μ on $\mathbb{R}^{d+1}$ satisfying the following conditions:

(I) There is a closed, convex set $\mathfrak{X} \subset \mathbb{R}^d$ with $\mathfrak{X}^\circ \neq \emptyset$ such that $\mu(\mathfrak{X} \times \mathbb{R}) = 1$. Also,

$$\int_{\mathfrak{X}\times\mathbb{R}} y^2 \mu(dx, dy) < \infty.$$

(II) There is a function $\phi : \mathbb{R}^d \to \mathbb{R}$ with $\mathfrak{X} \subset dom(\phi)$ such that whenever $(X, Y) \sim \mu$ we have $\mathbf{E}\left(Y - \phi(X)|X\right) = 0$ and $\mathbf{E}\left(|Y - \phi(X)|^2\right) = \sigma^2 < \infty$. Thus, ϕ is the regression function. Note that ϕ is not necessarily convex.

(III) Denoting by $\nu(\cdot) = \mu((\cdot) \times \mathbb{R})$ the x-marginal of μ, we assume that

$$\{\mathtt{X} \in \mathcal{B}_{\mathfrak{X}} : \nu(\mathtt{X}) = 0\} = \{\mathtt{X} \in \mathcal{B}_{\mathfrak{X}} : \mathtt{X} \text{ has Lebesgue measure } 0\}.$$

(IV)

$$\int_{\mathbb{R}^d} x^2 \nu(dx) < \infty.$$

Defining the set $\mathcal{K}_\nu$ by

$$\mathcal{K}_\nu := \left\{ \psi : \mathbb{R}^d \to \mathbb{R} \mid \psi \text{ is convex with } \int |\psi(x)|^2 \nu(dx) < \infty \right\},$$

it can be seen that $\mathcal{K}_\nu$ is a closed, convex cone in $\mathbb{L}^2(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d}, \nu)$. Consequently, from Moreau's decomposition theorem (see the proof of Lemma 2.2.4) there is a unique convex function $\hat{\phi} \in \mathcal{K}_\nu$ which satisfies the equations

$$\begin{aligned} \int_{\mathbb{R}^d} (\phi - \hat{\phi})^2 d\nu &= \inf_{\psi \in \mathcal{K}_\nu} \left\{ \int_{\mathbb{R}^d} (\phi - \psi)^2 d\nu \right\}; \\ \int_{\mathbb{R}^d} (\phi - \hat{\phi})(\psi - \hat{\phi}) d\nu &\leq 0 \quad \forall\, \psi \in \mathcal{K}_\nu. \end{aligned}$$

Since whenever $(X, Y) \sim \mu$ we have $\mathbf{E}\,(Y|X) = \phi(X)$, it is straightforward to see that $\hat{\phi}$ is the only function satisfying

$$\int_{\mathfrak{X} \times \mathbb{R}} (\psi(x) - \hat{\phi}(x))(y - \hat{\phi}(x)) \mu(dx, dy) \leq 0 \ \ \forall\, \psi \in \mathcal{K}_\nu.$$

From the arguments in Lim and Glynn (2012) it can be concluded that, under (I)-(IV), Theorems 2.3.1 and 2.3.2 hold with $\hat{\phi}$ replacing ϕ. Note that under {A5-A7} in Section 2.3 we obviously have $\hat{\phi} = \phi$. This is a remarkable property of the convex least squares estimator. Even if the regression model is misspecified by the researcher, the estimator has asymptotic regularity. It always converges to the Hilbert space projection of the actual regression function ϕ onto the closed convex cone $\mathcal{K}_\nu$ in $\mathbb{L}^2(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d}, \nu)$. The interested reader can look at all the details in Lim and Glynn (2012).

3.4 A conjecture about local rates of convergence

In this section we focus on the stochastic design regression model described in Section 2.3.2. Our aim is to state a conjecture about the *local* rate of convergence of the convex least squares estimator, namely, the rate at which $\hat{\phi}_n(x_0) - \phi(x_0)$ converges to 0 for a given $x_0 \in \mathcal{X}^\circ$. To do this, we will first introduce some additional notation and assumptions and then prove some facts about convex functions and empirical processes to provide the substance for the conjecture.

3.4.1 Some further notation and assumptions

We will now proceed to introduce some notation in addition to that introduced in Section 2.3. Consider a convex function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$. The closure of f is the greatest lower semicontinuous function majorized by f. A convex function is called closed if it equals its closure. The epigraph of f is the subset of $\mathbb{R}^{d+1}$ given by: $epi(f) := \{(x, y) \in \mathbb{R}^d \times \mathbb{R} = \mathbb{R}^{d+1} : y \geq f(x)\}$. Additionally, we will denote by f^* the *convex conjugate function* of f (or its *Legendre-Fenchel transform*). This function is given by

$$f^*(\xi) := \sup_{x \in \mathbb{R}^d} \{\langle \xi, x\rangle - f(x)\}.$$

The methods that will be used to back the conjecture of this section require an additional assumption on the convex regression function ϕ. We have to impose a condition stronger than mere convexity, which will be labeled as condition (A8).

(A8) There is a neighborhood $U \subset \mathfrak{X}$ of x_0 and two constants $C, c > 0$ such that the marginal density of X under μ, denoted by f_ν, and the function $x \mapsto \mu\left(|Y - \phi(X)|^2 | X = x\right)$ are continuous on U; ϕ and $\nabla\phi$ are continuously differentiable and Lipschitz on U, respectively; and

$$c|x_1 - x_2|^2 \ \leq \ \phi(x_2) - \phi(x_1) - \langle \nabla\phi(x_1), x_2 - x_1\rangle \ \leq \ C|x_1 - x_2|^2 \tag{3.5}$$

for all $x_1, x_2 \in U$.

Note that (3.5) will be satisfied whenever ϕ is twice continuously differentiable with nonsingular Hessian in some neighborhood of x_0. However, we do not assume second order differentiability.

3.4.2 On the measurability of subdifferentials and subgradients

In what follows we will provide some facts about set-valued measurable functions and random convex functions that will allow us to avoid the use of inner and outer probabilities.

It follows from Lemma 2.2.4 that $\hat{\phi}_n(x)$ is measurable for any $x \in \mathbb{R}^d$. According to the arguments in the second paragraph of page 6 in Rockafellar (1969), this statement, together with the fact that $\mathbf{P}\left(Conv\left(X_1, \ldots, X_n\right) \neq \emptyset\right) = 1$, implies that $\hat{\phi}_n$ is a *normal convex integrand* (see page 5 of Rockafellar (1969)). Consequently, the corresponding subdifferential mappings $(\partial\hat{\phi}_n)_{n=1}^{\infty}$ are *measurable multifunctions* (this follows from Corollary 4.6 in Rockafellar (1969)). More general notions of normal integrands and random lower semicontinuous functions can be found in Attouch and Wets (1990) and Hess (1995). For the basic properties of measurable set-valued mappings, the interested reader may look at Aubin and Frankowska (2009).

Given what has been stated in the previous paragraph, Corollary 8.2.13, page 317 in Aubin and Frankowska (2009) allows us to remove the inner probabilities in Theorems 2.3.1 and 2.3.2 (as the subdifferentials of the least squares estimators are measurable multifunctions). Moreover, the same result allows us to guarantee the measurability of the events and random variables considered in the sequel.

3.4.3 The conjecture and the ideas behind it

Conjecture: Under conditions {A5-A8} for $d = 1, 2, 3$ we have

(i) $\hat{\phi}_n(x_0) - \phi(x_0) = O_{\mathbf{P}}\left(n^{-\frac{2}{d+4}}\right)$;

(ii) $\sup_{\xi \in \partial \hat{\phi}_n(x_0)} \{|\xi - \nabla\phi(x_0)|\} = O_{\mathbf{P}}\left(n^{-\frac{1}{d+4}}\right).$

In what follows the main ideas to back up this claim will be provided. The remainder of this section constitutes the latest attempt to use these ideas to prove the conjecture. In the end, it will hopefully be clear what remains to be done to have a complete proof.

3.4.3.1 "Localizing" the least squares criterion

We will attempt to use in some clever way the "localizing" functions given in Lemmas A.2.1 and A.2.2 (see Section A.2). Suppose that we are given a compact, convex set $\mathtt{X} \subset U$ (where U is given in (A8)) and consider the functions $\underline{\hat{\phi}}_n$ and $\overline{\hat{\phi}}_n$ given by Lemmas A.2.1 and A.2.2, respectively. Define $\underline{\hat{\psi}}_n := \hat{\phi}_n - \underline{\hat{\phi}}_n$ and $\overline{\hat{\psi}}_n := \overline{\hat{\phi}}_n - \hat{\phi}_n$. Then, applying the characterization of the orthogonal projection on Euclidean spaces (see Lemma 2.4, (i) in Seijo and Sen (2011b)) these functions yield the inequalities:

$$\sum_{X_k \in \mathtt{X}} (Y_k - \hat{\phi}_n(X_k))\overline{\hat{\psi}}_n(X_k) \leq 0, \tag{3.6}$$

$$\sum_{X_k \in \mathtt{X}} (Y_k - \hat{\phi}_n(X_k))\underline{\hat{\psi}}_n(X_k) \geq 0. \tag{3.7}$$

Note that the lower bounds in Lemmas A.2.1 and A.2.2 imply that both of these functions are nonnegative. Thus, there is a continuous, nonnegative function $\hat{\psi}_n$ which is a convex combination of $\underline{\hat{\psi}}_n$ and $\overline{\hat{\psi}}_n$ (and thus supported on $\mathtt{X}$) such that

$$\sum_{X_k \in \mathtt{X}} (Y_k - \hat{\phi}_n(X_k))\hat{\psi}_n(X_k) = 0. \tag{3.8}$$

Before proceeding any further, I would like to compute these "localizing" functions in a particular example, so that we get a better idea of how they work. Let f be a quadratic form $f(x) = \langle (x - x_0), A(x - x_0)\rangle + \langle b, x - x_0\rangle + c$

for some symmetric positive definite matrix A and let $\mathtt{X} := \{x \in \mathbb{R}^d : \langle (x - x_0), A(x - x_0)\rangle \leq r\}$. Then, a straightforward calculation shows that $\overline{f}(x) - f(x) = r^2 - \langle (x - x_0), A(x - x_0)\rangle$ and $f(x) - \underline{f}(x) = (r - \langle (x - x_0), A(x - x_0)\rangle)^2$ for $x \in \mathtt{X}$. This example is particularly illustrative because if the regression function ϕ is twice continuously differentiable around x_0, the corresponding localizing functions will be close to these ones for $A = \frac{1}{2}\nabla^2\phi(x_0)$ and $\mathtt{X}$ will be similar to a ball around x_0 under the metric defined by the Hessian.

Coming back to the derivation of the rate, we can rewrite (3.8) as follows:

$$\sum_{X_k \in \mathtt{X}} (Y_k - \phi(X_k))\hat{\psi}_n(X_k) = \sum_{X_k \in \mathtt{X}} (\hat{\phi}_n(X_k) - \phi(X_k))\hat{\psi}_n(X_k). \tag{3.9}$$

The idea is to use this last identity to derive the rate. A description of how to do it will be presented in the following paragraphs.

For starters, consider $\alpha, r_0 > 0$ and $\rho \in (0, 1)$ such that (3.5) holds in $B(x_0, r_0)$ and $m_{\alpha,\rho} := \rho c - 4(\alpha + \sqrt{2\alpha(\alpha + C)}) > 0$. Define r_n as follows:

$$r_n := \inf\left\{r \geq 0 : \left\|\hat{\phi}_n - \phi\right\|_{B(x_0,s)} \leq \alpha s^2 \ \forall\ r \leq s \leq r_0\right\}. \tag{3.10}$$

Before elaborating further on the properties of r_n, let us introduce some additional notation. For any $r \leq r_0$ let $\underline{\hat{\phi}}_{n,r}$ and $\overline{\hat{\phi}}_{n,r}$ be the functions given by Lemmas A.2.1 and A.2.2, respectively, applied to $f = \hat{\phi}_n$ over $\mathcal{X} = \overline{B(x_0, r)}$. Define $\underline{\hat{\psi}}_{n,r} := \hat{\phi}_{n,r} - \underline{\hat{\phi}}_{n,r}$ and $\overline{\hat{\psi}}_{n,r} := \overline{\hat{\phi}}_{n,r} - \hat{\phi}_n$. As argued in the derivation of (3.9), there is $\lambda_{n,r} \in [0, 1]$ such that for $\hat{\psi}_{n,r} := \lambda_{n,r}\overline{\hat{\psi}}_{n,r} + (1 - \lambda_{n,r})\underline{\hat{\psi}}_{n,r} \geq 0$ we have:

$$\sum_{|x_0 - X_k| \leq r} (Y_k - \phi(X_k))\hat{\psi}_{n,r}(X_k) = \sum_{|x_0 - X_k| \leq r} (\hat{\phi}_n(X_k) - \phi(X_k))\hat{\psi}_{n,r}(X_k). \tag{3.11}$$

We will now summarize the main properties of r_n in the following Lemma:

Lemma 3.4.1 *Under $\{A5\text{-}A8\}$, the sequence of random variables* $(r_n)_{n=1}^{\infty}$ *converges to zero with probability one. Moreover, let* $n \in \mathbb{N}$ *and* $r_n \le r \le \frac{1}{2}r_0$. *Then,*

(i) $\|\partial\hat{\phi}_n - \nabla\phi\|_{B(x_0,r)} \le 2(\alpha + \sqrt{2\alpha(\alpha + C)})r.$

(ii) For all $w, y \in B(x_0, r)$ *we have*

$$\begin{aligned}\hat{\phi}_n(w) - \hat{\phi}_n(y) - \langle\partial\hat{\phi}_n(y), w - y\rangle &\ge |w - y|(c|w - y| - 4(\alpha + \sqrt{2\alpha(\alpha + C)})r)_+;\\ \hat{\phi}_n(w) - \hat{\phi}_n(y) - \langle\partial\hat{\phi}_n(y), w - y\rangle &\le |w - y|(C|w - y| + 4(\alpha + \sqrt{2\alpha(\alpha + C)})r);\end{aligned}$$

(iii) $\sup_{|w-x_0|\le r}\{\hat{\underline{\psi}}_{n,r}(w) \vee \hat{\overline{\psi}}_{n,r}(w)\} \le (C + 4(\alpha + \sqrt{2\alpha(\alpha + C)}))r^2.$

(iv) $\hat{\underline{\psi}}_{n,r}(w) \wedge \hat{\overline{\psi}}_{n,r}(w) \ge \frac{\rho}{2}m_{\alpha,\rho}r^2$ *for all* $w \in B(x_0, (1 - \rho)r)$

(v) $\|\hat{\overline{\phi}}_n - \phi\|_{B(x_0,r)} \vee \|\hat{\underline{\phi}}_n - \phi\|_{B(x_0,r)} \le (\alpha + C + 4(\alpha + \sqrt{2\alpha(\alpha + C)}))r^2.$

In addition, there is a constant $M_{C,\alpha}$ *such that whenever* $r_n \le r \le \frac{1}{4}r_0$ *we have,*

(vi) $\|\partial\hat{\overline{\phi}}_n - \nabla\phi\|_{B(x_0,r)} \vee \|\partial\hat{\underline{\phi}}_n - \nabla\phi\|_{B(x_0,r)} \le M_{C,\alpha}r$

Proof: The first assertion follows from the fact that $r_n \le \frac{1}{\sqrt{\alpha}}\sqrt{\|\hat{\phi}_n - \phi\|_{B(x_0,r_0)}} \xrightarrow{a.s.} 0$. Now, let $w \in B(x_0, r)$, $\delta > 0$ and $\partial\hat{\phi}_n(w)$ stand for any of its elements (recall that the subdifferential might be multivalued). Assume that $|\partial\hat{\phi}_n(w) - \nabla\phi(w)| \ne 0$. Define $y_\delta := w + \frac{r\delta}{|\partial\hat{\phi}_n(w) - \nabla\phi(w)|}(\partial\hat{\phi}_n(w) - \nabla\phi(w))$. Note then that by (A8), for all $\delta < r_0/r$ we have that $-\langle\nabla\phi(w), y_\delta - w\rangle \le -\phi(y) + \phi(w) + C|y_\delta - w|^2$. Using the defining property of the subdifferential (for $\partial\hat{\phi}_n(w)$) we get

$$\begin{aligned}r\delta|\partial\hat{\phi}_n(w) - \nabla\phi(w)| &= \langle\partial\hat{\phi}_n(w) - \nabla\phi(w), y_\delta - w\rangle\\ &\le (\hat{\phi}_n(y_\delta) - \phi(y_\delta)) - (\hat{\phi}_n(w) - \phi(w)) + C|y - w|^2\\ &\le \alpha r^2 + \alpha(1 + \delta)^2 r^2 + Cr^2\delta^2.\end{aligned}$$

Consequently, as $r_0/r \geq 2$,

$$\begin{aligned}|\partial\hat{\phi}_n(w) - \nabla\phi(w)| &\leq 2\alpha r + r \inf_{0<\delta\leq 2}\{\frac{2\alpha}{\delta} + (\alpha + C)\delta)\} \\ &\leq 2(\alpha + \sqrt{2\alpha(\alpha + C)})r.\end{aligned}$$

As w was arbitrarily chosen, this completes the proof of (i). Now, let $w, y \in B(x_0, r)$. From (i) and the mean value theorem of integral calculus we can write

$$\left|(\hat{\phi}_n(w) - \hat{\phi}_n(y) - \langle\partial\hat{\phi}_n(y), w - y\rangle) - (\phi(w) - \phi(y) - \langle\nabla\phi(y), w - y\rangle)\right| =$$

$$\left|\int_0^1 \langle(\partial\hat{\phi}_n - \nabla\phi)(y + t(w - y)), w - y\rangle dt + \langle(\partial\hat{\phi}_n - \nabla\phi)(y), w - y\rangle\right| \leq 4(\alpha+\sqrt{2\alpha(\alpha + C)})r|w-y|.$$

Putting the last inequality together with (A8) one immediately gets (ii) (via applications of the Cauchy-Schwartz and triangular inequalities). Furthermore, (iii) and (iv) are easily obtained from the bounds given in Lemma A.2.1 (iv) and Lemma A.2.2 (v). Finally, (v) follows immediately from (iii) and (vi) is proved like (i) with $(\alpha+C+4(\alpha+\sqrt{2\alpha(\alpha + C)}))$ and $\frac{1}{2}r_0$ assuming the roles of α and r_0, respectively. □

3.4.3.2 Some calculations involving empirical processes

For $r > 0$ consider the following class of functions:

$$\mathcal{C}^{a,b}_{x_0,r} := \left\{ f : B(x_0, r) \to \mathbb{R} \;\middle|\; \begin{array}{c} f \text{ is convex;} \\ \|f\|_{\overline{B(x_0,r)}} \leq ar^2; \\ \|\partial f\|_{\overline{B(x_0,r)}} \leq br \end{array} \right\},$$

where $\|\partial f\|_{\overline{B(x_0,r)}} = \sup\{|\xi| : \ \xi \in \partial f(x), \ |x - x_0| \leq r\}$. Additionally, we define the class

$$\mathcal{E}^{a,b}_{x_0,r} := (y-\phi(x))\mathcal{C}^{a,b}_{x_0,r} = \{h : \mathbb{R}^d\times\mathbb{R} \to \mathbb{R} : h(x, y) = (y-\phi(x))f(x) \text{ for some } f \in \mathcal{C}^{a,b}_{x_0,r}\}.$$

Note that $\mathcal{C}^{a,b}_{x_0,r}$ is a class of functions in $\mathbb{R}^d$, while $\mathcal{E}^{a,b}_{x_0,r}$ is a class of functions in $\mathbb{R}^{d+1}$ with measurable envelope $ar^2|y-\phi(x)|\mathbf{1}_{B(x_0,r)}(x)$.

We will now introduce some notation. We will denote by μ_n the empirical measure defined on $\mathbb{R}^{d+1}$ by the sample $(X_{n,1},Y_{n,1}),\dots,(X_{n,n},Y_{n,n})$. In agreement with van der Vaart and Wellner (1996), given a class of functions $\mathcal{G}$ on $D\subset\mathbb{R}^{d+1}$, a seminorm $\|\cdot\|$ on some space containing $\mathcal{G}$ and $\eta>0$ we denote by $N(\eta,\mathcal{G},\|\cdot\|)$ the η-covering number of $\mathcal{G}$ with respect to $\|\cdot\|$. Additionally, we will be making use of the notation $\|\gamma\|_{\mathcal{G}}$ for $\sup\left\{\int|f|d\gamma:\ f\in\mathcal{G}\right\}$ for any signed Borel measure γ on $\mathbb{R}^{d+1}$ (or some other Euclidean space) and any class $\mathcal{G}$ of measurable functions.

The next result presents entropy calculations for the classes $\mathcal{E}^{a,b}_{x_0,r}$ and $\mathcal{C}^{a,b}_{x_0,r}$.

Lemma 3.4.2 *Let $a,b,r>0$. Then, there is a constant $B_{a,b,d}>0$, depending only on a, b and d, such that*

(i) $\log\left(N(\eta,\mathcal{C}^{a,b}_{x_0,r},\|\cdot\|_{B(x_0,r)})\right)\le B_{a,b,d}r^d\eta^{-d/2}$,

(ii) $\sup_Q\left\{\log\left(N\left(\eta ar^2\|\epsilon\mathbf{1}_{B(x_0,r)}\|_{Q,2},\mathcal{E}^{a,b}_{x_0,r},\|\cdot\|_{Q,2}\right)\right)\right\}\le B_{a,b,d}a^{-d/2}\eta^{-d/2}$, *where* $\epsilon=y-\phi(x)$, $\|\cdot\|_{Q,2}$ *stands for the* $\mathbb{L}_2(Q)$ *norm and the supremum is taken over all discrete probability measures* Q *on* $\mathbb{R}^{d+1}$ *for which* $\|\epsilon\mathbf{1}_{B(x_0,r)}\|_{Q,2}>0$.

Proof: For each $f\in\mathcal{C}^{a,b}_{x_0,r}$, let $f_r:B(x_0,1)\to\mathbb{R}$ be given by $f_r(x)=r^{-2}f(rx)$. Then, $\|f_r\|_{B(x_0,1)}\le a$ and $\|\partial f_r\|_{B(x_0,1)}=r^{-1}\|\partial f\|_{B(x_0,r)}\le b$. It follows that $\{f_r\}_{f\in\mathcal{C}^{a,b}_{x_0,r}}$ is a subset of the space of all uniformly bounded, uniformly Lipschitz (i.e. with uniformly bounded Lipschitz constants) convex functions on $B(x_0,1)$. Moreover, $\|f-g\|_{B(x_0,r)}=r^2\|f_r-g_r\|_{B(x_0,1)}$. It follows that $r^{-2}\eta$-nets for $\{f_r\}_{f\in\mathcal{C}^{a,b}_{x_0,r}}$ have a one-to-one correspondence with η-nets for $\mathcal{C}^{a,b}_{x_0,r}$. Hence, (i) is a consequence of Theorem 6 in Bronšteĭn (1976) (see also Corollary 2.7.10 in page 164 of van der Vaart and Wellner (1996)). For (ii) note that if Q is any discrete probability measure and $f, g \in \mathcal{C}^{a,b}_{x_0,r}$ then $\|\epsilon f \mathbf{1}_{B(x_0,1)} - \epsilon g \mathbf{1}_{B(x_0,1)}\|_{Q,2} \le \|\epsilon \mathbf{1}_{B(x_0,r)}\|_{Q,2}\|f-g\|_{B(x_0,r)}$. It is now clear that (ii) follows from (i). $\square$

With the aid of Lemmas 3.4.1 and 3.4.2 we will attempt to derive a maximal inequality for certain classes $\mathcal{E}^{a,b}_{x_0,r}$. These inequalities will be expressed in terms of the following quantities defined for $\delta > 0$:

$$\begin{aligned}
\mathcal{J}^{a,b}_{x_0,r}(\delta) &:= \sup_Q \left\{ \int_0^\delta \sqrt{1 + \log\left(N\left(\eta a r^2 \|\epsilon^2 \mathbf{1}_{B(x_0,r)}\|_{Q,2}, \mathcal{E}^{a,b}_{x_0,r}, \|\cdot\|_{Q,2}\right)\right)} d\eta \right\};\\
\mathcal{I}_{a,b}(\delta) &:= \int_0^\delta \sqrt{1 + B_{a,b,d} a^{-d/2} \eta^{-d/2}} d\eta
\end{aligned}$$

where the supremum is again taken over all discrete probability measures Q on $\mathbb{R}^{d+1}$ for which $\|\epsilon^2 \mathbf{1}_{B(x_0,r)}\|_{Q,2} > 0$.

Lemma 3.4.3 *Let $a, b, r, M > 0$ and assume that $\{A5\text{-}A8\}$ hold. Assume that r is small enough so $B(x_0, r) \subset U$ and let $\varsigma_{\nu,r} := \sup\{f_\nu(x)\mu\left(|Y - \phi(X)|^2 | X = x\right) : |x - x_0| \le r\} < \infty$. Then,*

$$\mathbf{P}\left(\sup_{f \in \mathcal{C}^{a,b}_{x_0,r}} \left\{ \frac{1}{\sqrt{n}} \left| \sum_{k=1}^n \epsilon_k f(X_k) \right| \right\} > M \right) \le \frac{\varsigma_{\nu,r} a^2 \pi^{d/2} \mathcal{I}_{a,b}(1)}{M^2 \Gamma\left(\frac{d}{2}+1\right)} r^{d+4},$$

where the right-hand side of the preceding inequality is finite for $d = 1, 2, 3$.

Proof: Apply Theorem 2.14.1 in page 239 of van der Vaart and Wellner (1996) to the class $\mathcal{E}^{a,b}_{x_0,r}$ (note that the supremum inside the probability is measurable as $\mathcal{C}^{a,b}_{x_0,r}$ is separable under the sup-norm). In view of Lemma 3.4.2 we get:

$$\begin{aligned}
\mathbf{P}\left(\sup_{f \in \mathcal{C}^{a,b}_{x_0,r}} \left\{ \frac{1}{\sqrt{n}} \left| \sum_{k=1}^n \epsilon_k f(X_k) \right| \right\} > M \right) &\le \frac{1}{M^2}\mathbf{E}\left(\left(\sup_{f \in \mathcal{C}^{a,b}_{x_0,r}} \left\{ \frac{1}{\sqrt{n}} \left| \sum_{k=1}^n \epsilon_k f(X_k) \right| \right\}\right)^2\right)\\
&\le \mathcal{J}^{a,b}_{x_0,r}(1) \int_{|x-x_0|\le r} a^2 r^4 \mu\left(|Y-\phi(X)|^2 | X = x\right) f_\nu(x) dx\\
&\le \frac{\varsigma_{\nu,r} a^2 \pi^{d/2} \mathcal{I}_{a,b}(1)}{M^2 \Gamma\left(\frac{d}{2}+1\right)} r^{d+4}.
\end{aligned}$$

□

Now, we have the following result.

Lemma 3.4.4 *Let*

$$B_n := \frac{1}{n^{\frac{d+2}{d+4}} r_n^{d+2}} \sum_{|X_k - x_0| \le r_n} \epsilon_k \hat{\psi}_{n,r_n}(X_k),$$

and assume that {A5-A8} hold. Then, there is a constant $K_{\alpha,\rho,d}$, depending only on d, $\alpha > 0$ and $\rho \in (0,1)$ (the constants involved in in the definition of r_n), such that

$$\overline{\lim_{n\to\infty}} \mathbf{P}\left(|B_n| > b, r_n > an^{-\frac{1}{d+4}}\right) \le \frac{K_{\alpha,\rho,d}}{b^2} \sum_{k=1}^{\infty} \frac{k^{d+4}}{((k-1)\vee a)^{2d+4}}.$$

In particular, for $d \in \{2,3\}$, $\overline{\lim_{n\to\infty}} \mathbf{P}\left(|B_n| > b, r_n > an^{-\frac{1}{d+4}}\right) \to 0$ as $b \to \infty$ for any $a > 0$.

Proof: First note that condition (A8) implies that

$$\sup_{|x-x_0|\le r} \{|\phi(x) - \phi(x_0) - \langle \nabla\phi(x_0), x - x_0\rangle|\} \le Cr^2 \ \forall \ \ 0 \le r \le r_0,$$

and that $\nabla\phi$ is Lipschitz on U. Combining these two facts with Lemma 3.4.1 (i), (v) and (vi) there are two constants $a_{\alpha,\rho}, b_{\alpha,\rho} > 0$, which only depend on the $\rho, \alpha > 0$ involved in the definition of r_n, such that

$$\begin{aligned}
\overline{\hat{\phi}}_{n,s}(\cdot) - \phi(x_0) - \langle \nabla\phi(x_0), (\cdot) - x_0\rangle &\in \mathcal{C}_{x_0,r}^{a_{\alpha,\rho},b_{\alpha,\rho}};\\
\underline{\hat{\phi}}_{n,s}(\cdot) - \phi(x_0) - \langle \nabla\phi(x_0), (\cdot) - x_0\rangle &\in \mathcal{C}_{x_0,r}^{a_{\alpha,\rho},b_{\alpha,\rho}};\\
\hat{\phi}_{n,s}(\cdot) - \phi(x_0) - \langle \nabla\phi(x_0), (\cdot) - x_0\rangle &\in \mathcal{C}_{x_0,r}^{a_{\alpha,\rho},b_{\alpha,\rho}} \ \forall \ r_n \le s \le r \le \frac{1}{4} r_0.
\end{aligned} \tag{3.12}$$

Let $m_n := \lfloor \frac{r_0}{4} n^{\frac{1}{d+4}} \rfloor$. Then we have

$$\mathbf{P}\left(|B_n| > b, r_n > an^{-\frac{1}{d+4}}\right) \le \mathbf{P}\left(r_n > m_n n^{\frac{1}{d+4}}\right)$$

$$+ \sum_{kn^{\frac{1}{d+4}} \le \frac{r_0}{2}} \mathbf{P}\left(\frac{1}{\sqrt{n}} \left| \sum_{|X_k - x_0| \le r_n} \epsilon_k \hat{\psi}_{n,r_n}(X_k) \right| > bn^{\frac{d}{2(d+4)}} r_n^{d+2}, ((k-1) \vee a) n^{-\frac{1}{d+4}} < r_n \le kn^{-\frac{1}{d+4}} \right).$$

Note that $\hat{\psi}_{n,r_n}$ can be written as follows:

$$\begin{aligned} \hat{\psi}_{n,r_n} &= \lambda_{n,r_n}(\overline{\hat{\phi}}_{n,s}(\cdot) - \phi(x_0) - \langle \nabla \phi(x_0), (\cdot) - x_0 \rangle) \\ &- (1 - \lambda_{n,r_n})(\underline{\hat{\phi}}_{n,s}(\cdot) - \phi(x_0) - \langle \nabla \phi(x_0), (\cdot) - x_0 \rangle) \\ &+ (1 - 2\lambda_{n,r})(\hat{\phi}_{n,s}(\cdot) - \phi(x_0) - \langle \nabla \phi(x_0), (\cdot) - x_0 \rangle). \end{aligned}$$

Considering (3.12), Lemma 3.4.3 and the last inequality we get:

$$\mathbf{P}\left(|B_n| > b, r_n > an^{-\frac{1}{d+4}} \right) \le \mathbf{P}\left(r_n > m_n n^{\frac{1}{d+4}} \right)$$

$$+ \sum_{kn^{\frac{1}{d+4}} \le \frac{r_0}{4}} \mathbf{P}\left(\sup_{f \in \mathcal{C}^{a_{\alpha,\rho}, b_{\alpha,\rho}}_{x_0, kn^{-1/(d+4)}}} \left\{ \frac{1}{\sqrt{n}} \left| \sum_{k=1}^{n} \epsilon_k f(X_k) \right| \right\} > \frac{b((k-1) \vee a)^{d+2} n^{-1/2}}{3} \right)$$

$$\le \mathbf{P}\left(r_n > m_n n^{\frac{1}{d+4}} \right) + \sum_{kn^{\frac{1}{d+4}} \le \frac{r_0}{4}} \frac{\varsigma_{\nu,r} a^2_{\alpha,\rho} \pi^{d/2} \mathcal{I}_{a_{\alpha,\rho}, b_{\alpha,\rho}}(1)}{b^2((k-1) \vee a)^{2d+4} \Gamma\left(\frac{d}{2} + 1\right)} k^{d+4}.$$

The result now easily follows, as from Lemma 3.4.1 we know that $r_n \xrightarrow{a.s.} 0$.

□

3.4.3.3 The consequences of Lemma 3.4.4 and what remains to be done

If we write $V_n := \frac{1}{nr_n^{d+2}} \sum_{k=1}^{n} \hat{\psi}_{n,r_n}(X_k)$ then it is easily seen (via a Glivenko-Cantelli type of argument) that there are constants $\beta_1, \beta_2 > 0$ such that $\mathbf{P}(\beta_1 < V_n < \beta_2) \to 1$ as $n \to \infty$. Moreover, if we let

$$U_n := \int_0^1 \left(\frac{1}{nr_n^{d+4}} \sum_{k=1}^{n} \langle (\partial \hat{\phi}_n - \nabla \phi)(x_0 + t(X_k - x_0)), X_k - x_0 \rangle \hat{\psi}_{n,r_n}(X_k) \right) dt$$

then it is also easily shown that $U_n = O_{\mathbf{P}}(1)$. Thus, Lemma 3.4.4 and (3.11) imply that we have (on the set where $r_n > an^{-\frac{1}{d+4}}$),

$$n^{\frac{2}{d+4}}(\hat{\phi}_n(x_0) - \phi(x_0))V_n + n^{\frac{2}{d+4}} r_n^2 U_n = O_{\mathbf{P}}(1). \tag{3.13}$$

As $\mathbf{P}\left(\beta_1 < V_n < \beta_2\right) \to 1$ as $n \to \infty$, to obtain the rate for $\hat{\phi}_n(x_0) - \phi(x_0)$ it suffices to show that $n^{\frac{2}{d+4}} r_n^2 U_n = O_{\mathbf{P}}(1)$. Moreover, if we were able to obtain an asymptotic lower bound for $|U_n|$ (just like β_1 for V_n), we would actually get a rate for r_n.

Part II

Change-point Regression

Chapter 4

A continuous mapping theorem for the smallest argmax functional

4.1 Introduction

Many estimators in statistics are defined as the maximizers of certain stochastic processes, called objective functions. This procedure for computing estimators is known as M-estimation and is quite common in modern statistics. A standard way to find the asymptotic distribution of a given M-estimator, is to obtain the limiting law of the (appropriately normalized) objective function and then apply the so-called argmax continuous mapping theorem (see Theorem 3.2.2, page 286 of van der Vaart and Wellner (1996) for a quite general version of this result). Chapter 3.2 in van der Vaart and Wellner (1996) gives an excellent account of M-estimation problems and applications of the argmax continuous mapping theorem.

Despite its proven usefulness in a wide range of applications, there are some M-estimation problems that cannot be solved by an application of the usual argmax continuous mapping theorem. This is particularly true when the objective functions converge in distribution to the law of some process that admits multiple maximizers. This situation arises frequently in problems concerning change-point estimation in regression settings. In these problems, the estimators are usually maximizers of processes that converge in the limit to two-sided, compound Poisson processes that have a complete interval of maximizers. See, for instance, Kosorok (2008b) (Section 14.5.1, pages 271–277), Lan et al. (2009), Kosorok and Song (2007), Pons (2003) and Seijo and Sen (2011a). This issue has been noted before by several authors, such as Ferger (2004).

The main goal of this chapter is to derive a version of the argmax continuous mapping theorem specially tailored for situations like the one described in the previous paragraph. A distinctive feature of the argmax continuous mapping theorem in this setup is that it requires the weak convergence, not only of the objective functions, but also of some associated *pure jump processes*. Although this requirement has been overlooked by some authors in the past (we discuss these omissions in Section 4.5), its necessity can be easily shown; see Section 4.4 for an example.

To illustrate the situations on which our results are applicable, we start with the following simple problem that arises in least squares change-point regression. Detailed accounts of this type of models can be found in Kosorok (2008b) (Section 14.5.1, pages 271–277), Lan et al. (2009) and Seijo and Sen (2011a). In its simplest form the model considers a random vector $X = (Y, Z)$ satisfying the following relation:

$$Y = \alpha_0 \mathbf{1}_{Z \le \zeta_0} + \beta_0 \mathbf{1}_{Z > \zeta_0} + \epsilon, \tag{4.1}$$

where Z is a continuous random variable, $\alpha_0 \neq \beta_0 \in \mathbb{R}$, $\zeta_0 \in [c_1, c_2] \subset \mathbb{R}$ and

ϵ is a continuous random variable, independent of Z with zero expectation and finite variance $\sigma^2 > 0$. The parameter of interest is ζ_0, the change-point. Given a random sample from this model, the *least squares estimator* $\hat{\theta}_n$ of $\theta_0 = (\zeta_0, \alpha_0, \beta_0) \in \Theta := [c_1, c_2] \times \mathbb{R}^2$ is obtained by maximizing the criterion function

$$M_n(\theta) := -\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \alpha\mathbf{1}_{Z_i \leq \zeta} + \beta\mathbf{1}_{Z_i > \zeta}\right)^2,$$

i.e.,

$$\hat{\theta}_n := (\hat{\zeta}_n, \hat{\alpha}_n, \hat{\beta}_n) = \operatorname*{sargmax}_{\theta\in\Theta}\{M_n(\theta)\}, \tag{4.2}$$

where sargmax denotes the maximizer with the smallest ζ value. This distinction is made as there is no unique maximizer for ζ, in fact, for any α, β, $M_n(\cdot, \alpha, \beta)$ is constant on every interval $[Z_{(j)}, Z_{(j+1)})$, where $Z_{(j)}$ stands for the j-th order statistic. It can be shown, see either Kosorok (2008b) (Section 14.5.1, pages 271–277) or Seijo and Sen (2011a), that $n(\hat{\zeta}_n - \zeta_0)$ converges in distribution to the smallest maximizer a two-sided, compound Poisson process. The convergence results in this chapter, Theorems 4.3.1 and 4.3.2, can, in particular, be applied to derive the asymptotic distribution of this estimator (see Sections 4.5.1 and 5.3).

Our results will be applicable to M-estimation problems for which the objective function takes arguments in some compact rectangle $K \subset \mathbb{R}^d$, $d \geq 1$. We focus on functions belonging to the Skorohod space $\mathcal{D}_K$ as defined in Neuhaus (1971). The elements of $\mathcal{D}_K$ are functions with finite "quadrant limits" (generalized one-sided limits) and are "continuous from above" (generalization of right-continuity) at each point in K. In Section 4.2 we describe the Skorohod space $\mathcal{D}_K$ in details and state some fundamental properties of the sargmax functional. Some of the results developed in this connection can also be of independent interest. In Section 4.3 we prove a version of the continuous mapping theorem for the sargmax functional for elements of $\mathcal{D}_K$ which are cádlág in the first component and jointly continuous on the last $d-1$. In Section 4.4 we describe an example that illustrates the necessity of the convergence of the associated pure jump processes in the results of Section 4.3. Finally, in Section 4.5 we apply the theorems of Section 4.3 to the change-point regression problem described above and to the estimation of a change-point in time and in a covariate in the Cox-proportional hazards model.

4.2 The Skorohod space $\mathcal{D}_K$

4.2.1 Definition and basic properties

We start by recalling the Skorohod space as discussed in Neuhaus (1971). To simplify notation, we write the coordinates of any vector in $\mathbb{R}^d$ with upper indices. We consider a compact rectangle $K = [a, b] = [a^1, b^1] \times \cdots \times [a^d, b^d]$ for some $a < b \in \mathbb{R}^d$ with the inequality holding componentwise. For any space $\mathbb{R}^m$ we will write $|\cdot|$ for the Euclidian norm (although the $\mathbb{L}^\infty$-norm is used in Neuhaus (1971), the results in there hold if one uses the Euclidian norm instead). For $k \in \{1, \ldots, d\}$, $t \in [a^k, b^k]$ and $s \in \{a^k, b^k\}$ we write:

$$
\begin{aligned}
I_k(s,t) &:= \begin{cases} [a^k, t) & \text{if } s = a^k, \\ (t, b^k] & \text{if } s = b^k. \end{cases} \\
J_k(s,t) &:= \begin{cases} [a^k, t) & \text{if } s = a^k \text{ and } t < b^k, \\ \left[a^k, b^k\right] & \text{if } s = a^k \text{ and } t = b^k, \\ \emptyset & \text{if } s = b^k \text{ and } t = b^k, \\ \left[t, b^k\right] & \text{if } s = b^k \text{ and } t < b^k. \end{cases}
\end{aligned}
$$

and for any $\rho \in \mathcal{V} := \prod_{k=1}^{d} \{a^k, b^k\}$, $x = (x^1, \ldots, x^d) \in \mathbb{R}^d$,

$$\begin{aligned} Q(\rho, x) &:= \prod_{k=1}^{d} I_k(\rho^k, x^k), \\ \tilde{Q}(\rho, x) &:= \prod_{k=1}^{d} J_k(\rho^k, x^k). \end{aligned}$$

Remark: Some properties of the sets $\tilde{Q}(\rho, x)$ are:

(a) $\tilde{Q}(\rho, x) \cap \tilde{Q}(\gamma, x) = \emptyset$ for every $\gamma \neq \rho \in \mathcal{V}$ and every $x \in K$.

(b) $K = \bigcup_{\rho \in \mathcal{V}} \tilde{Q}(\rho, x)$ for every $x \in K$.

Hence, $\left\{\tilde{Q}(\rho, x)\right\}_{\rho \in \mathcal{V}}$ forms a partition of K. We are now in a position to define the so-called quadrant limits, the concept of continuity from above and the Skorohod space.

Definition 4.2.1 (Quadrant Limits and Continuity from Above) *Consider a function $f : \mathbb{R}^d \to \mathbb{R}$, $\rho \in \mathcal{V}$ and $x \in K$. We say that a number l is the ρ-limit of f at x if for every sequence $\{x_n\}_{n=1}^{\infty} \subset Q(\rho, x)$ satisfying $x_n \to x$ we have $f(x_n) \to l$. In this case we write $l = f(x + 0_\rho)$. When $\rho = b$ we may write $f(x + 0_+) := f(x + 0_b)$. With this notation, f is said to be continuous from above at x if $f(x + 0_+) = f(x)$.*

Definition 4.2.2 (The Skorohod Space) *We define the Skorohod space $\mathcal{D}_K$ as the collection of all functions $f : K \to \mathbb{R}$ which have all ρ-limits and are continuous from above at every $x \in K$.*

Remark: It is easily seen that if $f \in \mathcal{D}_K$, $\rho \in \mathcal{V}$, $x \in K$ and $\{x_n\}_{n=1}^{\infty} \subset \tilde{Q}(\rho, x)$ is a sequence with $x_n \to x$, then $f(x_n) \to f(x + 0_\rho)$. This follows from the continuity from above as $Q(\rho, x) \cap Q(b, \xi) \neq \emptyset$ for every $\xi \in \tilde{Q}(\rho, x)$.

Before stating some of the most important properties of $\mathcal{D}_K$ we will introduce some further notation. Consider the partitions $\mathcal{T}_j = \{a^j = t_{j,0} < t_{j,1} < \ldots < t_{j,r_j} = b^j\}$ for $j = 1, \ldots, d$. We define the rectangular partition $\mathcal{R}(\mathcal{T}_1, \ldots, \mathcal{T}_d)$ determined by $\mathcal{T}_1, \ldots, \mathcal{T}_d$ as the collection of all rectangles of the form

$$R = \prod_{k=1}^{d} [t_{k,j_k-1}, t_{k,j_k}\rangle \ , j_k \in \{1, \ldots, r_k\}, \ k = 1, \ldots, d,$$

where $\rangle$ stands for ")" or "]" if $t_{k,j_k} < b^k$ or $t_{k,j_k} = b^k$, respectively. With the aid of this notation, we can now state two important lemmas.

Lemma 4.2.1 *Let $f \in \mathcal{D}_K$. Then, for every $\epsilon > 0$ there is $\delta > 0$ and partitions $\mathcal{T}_j$ of $[a^j, b^j]$, $j = 1, \ldots, d$, such that for any $R \in \mathcal{R}(\mathcal{T}_1, \ldots, \mathcal{T}_d)$ and any $\theta, \vartheta \in R$ with $|\theta - \vartheta| < \delta$ the inequality $|f(\theta) - f(\vartheta)| < \epsilon$ holds. Furthermore, we can take the partitions in such a way that $\sup_{\theta,\vartheta \in R} \{|\theta - \vartheta|\} < \delta$ for every $R \in \mathcal{R}(\mathcal{T}_1, \ldots, \mathcal{T}_d)$.*

Lemma 4.2.2 *Every function in $\mathcal{D}_K$ is bounded on K.*

Lemmas 4.2.1 and 4.2.2 are, respectively, Lemma 1.5 and Corollary 1.6 in Neuhaus (1971). Their proofs can be found there.

Let $K_1 = [a^1, b^1]$ and $K_2 = [a^2, b^2] \times \cdots \times [a^d, b^d]$, so $K = K_1 \times K_2$. We will be dealing with functions which are cádlág on the first coordinate and continuous on the remaining $d-1$. For this purpose we will turn our attention to the space $\widetilde{\mathcal{D}}_K \subset \mathcal{D}_K$ of all functions $f \in \mathcal{D}_K$ such that $f(t, \cdot) : K_2 \to \mathbb{R}$ is continuous $\forall\, t \in K_1$ and $f(\cdot, \xi) : K_1 \to \mathbb{R}$ is cádlág $\forall\, \xi \in K_2$.

Remark: It is worth noting that *all elements in $\mathcal{D}_K$ are componentwise cádlág*, so it is really the continuity in the last $d-1$ coordinates what makes $\widetilde{\mathcal{D}}_K$ a proper subspace of $\mathcal{D}_K$.

Lemma 4.2.3 *Let $f \in \widetilde{\mathcal{D}}_K$ and $\epsilon > 0$. Then, there is $\delta > 0$ such that*

$$\sup_{\substack{|\xi-\eta|<\delta \\ \xi,\eta\in K_2}} \{|f(t,\xi) - f(t,\eta)|\} \leq \epsilon \quad \forall\ t \in K_1.$$

Proof: From Lemma 4.2.1 we can find $\delta_0 > 0$ and partitions $\mathcal{T}_j$ of $[a^j, b^j]$, $j = 1, \ldots, d$ such that the conclusions of the lemma hold true with ϵ replaced by $\frac{\epsilon}{3}$. We take the partitions in such a way that whenever θ and ϑ belong to the same rectangle, the distance between them is less than δ_0. Let $s \in \mathcal{T}_1$. Since K_2 is compact and $f(s,\cdot)$ is continuous, we can find δ_s such that for any $\xi, \eta \in K_2$ with $|\xi - \eta| < \delta_s$ we get $|f(s,\xi) - f(s,\eta)| < \frac{\epsilon}{3}$. Let $\delta = \min_{s\in\mathcal{T}_1}\{\delta_s\}$ and pick $t \in K_1$ and $\xi, \eta \in K_2$ with $|\xi - \eta| < \delta$. Take the largest $s \in \mathcal{T}_1$ with $s \leq t$. Then, $|s - t| < \delta_0$ and hence

$$|f(t,\eta)-f(t,\xi)| \leq |f(t,\xi)-f(s,\xi)|+|f(s,\eta)-f(s,\xi)|+|f(t,\eta)-f(s,\eta)| < \epsilon.$$

The proof is then finished by taking the supremum over ξ and η and noticing that the choice of δ was independent of t. □

4.2.2 The Skorohod topology

So far we have not yet defined a topology on $\mathcal{D}_K$, so we turn our attention to this issue now. We will start by defining the Skorohod metric as given in Neuhaus (1971). Then, we will define a second metric on $\widetilde{D}_K$ and show that it is equivalent to the corresponding restriction of the Skorohod metric. This second metric will be more natural for the structure of $\widetilde{D}_K$ and will prove useful in the proof of the continuous mapping theorem for the smallest argmax functional. In order to define both of these metrics and state some of their properties, we will need some additional notation.

Consider a closed interval $I \subset \mathbb{R}$ and the class Λ_I of all functions $\lambda : I \to I$ which are surjective (onto) and strictly monotone increasing. Define the function $|||\cdot|||_I : \Lambda_I \to \mathbb{R}$ by the formula $|||\lambda|||_I = \sup_{s\neq t}\left\{\left|\log\left(\frac{\lambda(t)-\lambda(s)}{t-s}\right)\right|\right\}$. We write $\Lambda_K := \Lambda_{[a^1,b^1]}\times\cdots\times\Lambda_{[a^d,b^d]}$ and for $\lambda := (\lambda_1,\dots,\lambda_d) \in \Lambda_K$, $|||\lambda|||_K := \max_{1\leq k\leq d}\{|||\lambda_k|||_{[a^k,b^k]}\}$. In a similar fashion, we define $\Lambda_{K_2} := \Lambda_{[a^2,b^2]}\times\cdots\times\Lambda_{[a^d,b^d]}$ and for $\lambda \in \Lambda_{K_2}$, $|||\lambda|||_{K_2} := \max_{2\leq k\leq d}\{|||\lambda_k|||_{[a^k,b^k]}\}$. Note that for $(\lambda_1,\lambda) \in \Lambda_K = \Lambda_{K_1}\times\Lambda_{K_2}$ we have $|||(\lambda_1,\lambda)|||_K = |||\lambda_1|||_{K_1}\vee|||\lambda|||_{K_2}$. We will use the sup-norm notation also: for a function $f : A \to \mathbb{R}$ we write $\|f\|_A = \sup_{x\in A}\{|f(x)|\}$.

Definition 4.2.3 (The Skorohod metric) *We define the Skorohod metric $d_K : \mathcal{D}_K \times \mathcal{D}_K \to \mathbb{R}$ as follows:*

$$d_K(f,g) = \inf_{\lambda\in\Lambda_K}\{|||\lambda\,|||_K + \|f - g\circ\lambda\|_K\}\,.$$

With this definition we can now state the following fundamental result about the Skorohod space.

Lemma 4.2.4 *The Skorohod metric is a metric. If $\mathcal{D}_K$ is endowed with the topology defined by d_K, then it becomes a Polish space.*

For a proof of the last result, we refer the reader to Section 2 in Neuhaus (1971). We now proceed to define another metric, $\widetilde{d}_K$, on $\mathcal{D}_K$ by the formula:

$$\widetilde{d}_K(f,g) = \inf_{\lambda\in\Lambda_{[a^1,b^1]}}\left\{|||\lambda\,|||_{[a^1,b^1]} + \sup_{(t,\xi)\in K_1\times K_2}\{|f(t,\xi) - g(\lambda(t),\xi)|\}\right\}.$$

To properly describe the properties of $\widetilde{d}_K$ we need the ball notation for metric spaces: given a metric space $(\mathtt{X}, d)$, $r > 0$ and $x \in \mathtt{X}$ we write $B^d_r(x)$ for the open ball of radius r and center at x with respect to the metric d. Additionally, the following lemma will prove to be useful.

Lemma 4.2.5 *Let $I \subset \mathbb{R}$ be any compact interval. Then, for $\epsilon > 0$ there is $\delta > 0$ such that for any $\lambda \in \Lambda_I$ with $|||\lambda|||_I < \delta$ we also have*

$$\sup_{s \in I}\{|\lambda(s) - s|\} < \epsilon.$$

Proof: Assume that $I = [u, v]$. It suffices to choose $\delta < \frac{1}{4} \wedge \frac{\epsilon}{2|v-u|}$. To see this, observe that for any $\tau \in (0, \frac{1}{4})$, $\tau < 2\tau - 4\tau^2 \leq \log(1 + 2\tau)$ and for any $\tau > -1$, $\log(1 + \tau) \leq \tau$. It follows that for $\lambda \in \Lambda_I$ with $|||\lambda|||_I < \delta$ and any $s \in I$, $\log(1 - 2\delta) < -\delta \leq \log \frac{\lambda(s)-u}{s-u} \leq \delta < 2\delta - 4\delta^2 \leq \log(1 + 2\delta)$ and thus, $|\lambda(s) - s| < 2(s - u)\delta \leq 2|u - v|\delta$. In the previous inequalities we have made implicit use of the fact that $\lambda(u) = u$. □

The next lemma contains some of the most relevant properties of $\widetilde{d}_K$.

Lemma 4.2.6 *The following statements are true:*

(i) $\widetilde{d}_K$ *is a metric on* $\mathcal{D}_K$.

(ii) $d_K(f, g) \leq \widetilde{d}_K(f, g) \leq \|f - g\|_K \;\forall\; f, g \in \mathcal{D}_K$.

(iii) If $f \in \widetilde{\mathcal{D}}_K$, *then for every* $r > 0$ *there is* $\delta > 0$ *such that* $B^{d_K}_\delta(f) \subset B^{\widetilde{d}_K}_r(f)$. *Moreover, the metrics* d_K *and* $\widetilde{d}_K$ *generate the same topology on* $\widetilde{\mathcal{D}}_K$.

(iv) If f *is continuous, then for every* $r > 0$ *there is* $\delta > 0$ *such that* $B^{\widetilde{d}_K}_\delta(f) \subset B^{\|\cdot\|_K}_r(f)$. *Moreover, the metrics* d_K *and* $\widetilde{d}_K$ *and* $\|\cdot\|_K$ *generate the same topology on the space of continuous functions on* K.

(v) $(\widetilde{\mathcal{D}}_K, \widetilde{d}_K)$ *is a Polish space.*

Proof: It is straightforward to see that (ii) holds. The proof of (i) follows along the lines of the proof of the analogous results for the classical Skorohod metric (see Chapter 3 of Billingsley (1968)). For the sake of brevity we omit these arguments. For (iii) we use Lemma 4.2.3. Let $f \in \widetilde{\mathcal{D}}_K$, $r > 0$ and take $\delta_1 > 0$ such that the conclusions of Lemma 4.2.3 hold with $\frac{r}{3}$ replacing ϵ. Also, consider $\delta_2 > 0$ such that $|||\lambda|||_{K_2} < \delta_2$ implies $\sup_{\xi \in K_2}\{|\lambda(\xi) - \xi|\} < \delta_1$ (whose existence is a consequence of Lemma 4.2.5 applied to each of the intervals $[a^2, b^2], \ldots, [a^d, b^d]$). Let $\delta = \delta_2 \wedge \frac{r}{3}$ and take $g \in B_\delta^{d_K}(f)$. Find $(\lambda_1, \lambda) \in \Lambda_K = \Lambda_{K_1} \times \Lambda_{K_2}$ such that $|||(\lambda_1, \lambda)|||_K < \delta$ and $\|g - f \circ (\lambda_1, \lambda)\|_K < \frac{r}{3}$. Then, for any $(t, \xi) \in K_1 \times K_2$ we have:

$$\begin{aligned} |g(t,\xi) - f(\lambda_1(t), \xi)| &\leq |g(t,\xi) - f(\lambda_1(t), \lambda(\xi))| + |f(\lambda_1(t), \lambda(\xi)) - f(\lambda_1(t), \xi)| \\ &< \frac{r}{3} + \frac{r}{3}, \end{aligned}$$

where the second term in the sum of the right-hand side of the first inequality in the preceding display is less than $\frac{r}{3}$ because of Lemma 4.2.3 since $|||\lambda|||_{K_2} < \delta_2$. Taking supremum over $(t, \xi) \in K$ and considering that $|||\lambda_1|||_{K_1} < \frac{r}{3}$ we get that $\widetilde{d}_K(f, g) < r$. Thus, $B_\delta^{d_K}(f) \subset B_r^{\widetilde{d}_K}(f)$. Taking (ii) into account we can conclude that $\widetilde{d}_K$ and d_K are equivalent metrics on $\widetilde{\mathcal{D}}_K$.

We now turn out attention to (iv). Let $r > 0$. Then, there is $\delta_1 > 0$ such that $|f(x) - f(y)| < \frac{r}{2}$ whenever $|x - y| < \delta_1$. Also, there is $\delta_2 > 0$ such that $|||\lambda|||_{K_1} < \delta_2$ implies $\sup_{t \in K_1}\{|\lambda(t) - t|\} < \delta_1$. Let $\delta = \delta_2 \wedge \frac{r}{2}$ and let $g \in \mathcal{D}_K$ with $\widetilde{d}_K(f, g) < \delta$ and $\lambda \in \Lambda_{K_1}$ such that $|||\lambda\,|||_{K_1} + \|g(\cdot,\cdot) - f(\lambda(\cdot), \cdot)\|_{K_1 \times K_2} < \delta$. Then, for any $(t, \xi) \in K_1 \times K_2$ we have

$$|f(t,\xi) - g(t,\xi)| \leq |f(t,\xi) - f(\lambda(t), \xi)| + |f(\lambda(t), \xi) - g(t, \xi)| < r.$$

Thus, $B_\delta^{\widetilde{d}_K}(f) \subset B_r^{\|\cdot\|_K}(f)$.

To prove (v) it suffices to show that $\widetilde{\mathcal{D}}_K$ is a closed subset of $\mathcal{D}_K$, as the latter space is known to be Polish (see Neuhaus (1971)). Let $(f_n)_{n=1}^\infty$ be a sequence in $\widetilde{\mathcal{D}}_K$ such that $f_n \xrightarrow{d_K} f$ for some $f \in \mathcal{D}_K$. We will show that $f(t, \cdot)$ is continuous for every t and that will imply that $f \in \widetilde{\mathcal{D}}_K$ since f is automatically componentwise cádlág. Let $(t,\xi) \in K_1 \times K_2 = K$ and $\epsilon > 0$. Consider $n \in \mathbb{N}$ large enough so that $d_K(f, f_n) < \frac{\epsilon}{3}$ and take $\delta_1 > 0$ such that the conclusions of Lemma 4.2.3 hold true for f_n and $\frac{\epsilon}{3}$. Let $(\lambda_{n,1}, \lambda_n) \in \Lambda_{K_1} \times \Lambda_{K_2}$ such that $|||(\lambda_{n,1}, \lambda_n)\,|||_K + \|f - f_n \circ (\lambda_{n,1}, \lambda_n)\|_K < \frac{\epsilon}{3}$. Since λ_n is continuous, there is $\delta > 0$ such that $|\xi - \eta| < \delta$ implies $|\lambda_n(\xi) - \lambda_n(\eta)| < \delta_1$. It follows that $|f_n(\lambda_{n,1}(t), \lambda_n(\xi)) - f_n(\lambda_{n,1}(t), \lambda_n(\eta))| < \frac{\epsilon}{3}$ whenever $|\xi - \eta| < \delta$. Hence,

$$
\begin{aligned}
|f(t,\xi) - f(t,\eta)| &\le |f(t,\xi) - f_n(\lambda_{n,1}(t), \lambda_n(\xi))| + |f(t,\eta) - f_n(\lambda_{n,1}(t), \lambda_n(\eta))| \\
&\quad + |f_n(\lambda_{n,1}(t), \lambda_n(\xi)) - f_n(\lambda_{n,1}(t), \lambda_n(\eta))| \\
&< \epsilon, \quad \forall\, \xi, \eta \in K_2 \text{ such that } |\xi - \eta| < \delta.
\end{aligned}
$$

It follows that $f(t,\cdot)$ is continuous for every $t \in K_1$. Hence, $f \in \widetilde{\mathcal{D}}_K$ and $\widetilde{\mathcal{D}}_K$ is closed. $\square$

Remark: Observe that the previous lemma implies that for a convergent sequence in $\mathcal{D}_K$ with a limit in $\widetilde{\mathcal{D}}_K$ convergence in the $\widetilde{d}_K$ and d_K metrics are equivalent. When the limit is continuous, convergence in any of these metrics is equivalent to convergence in the sup-norm topology.

4.2.3 The sargmax functional on $\mathcal{D}_K$

We now turn our attention to the smallest argmax functional on $\mathcal{D}_K$.

Definition 4.2.4 (The sargmax Functional) *A function $f \in \mathcal{D}_K$ is said to have a maximizer at a point $x \in K$ if any of the quadrant-limits of x equals $\sup\limits_{\xi \in K}\{f(\xi)\}$. For any $f \in \mathcal{D}_K$ we can define the smallest argmax of f over the compact rectangle K, denoted by $\underset{x \in K}{\mathrm{sargmax}}\{f(x)\}$, as the unique element $x = (x^1, \dots, x^d) \in K$ satisfying the following properties:*

(i) x *is a maximizer of* f *over* K*,*

(ii) *if* $\xi = (\xi^1, \ldots, \xi^d)$ *is any other maximizer, then* $x^1 \le \xi^1$*,*

(iii) *if* ξ *is any maximizer satisfying* $x^j = \xi^j \; \forall \; j = 1, \ldots, k$ *for some* $k \in \{1, \ldots, d-1\}$*, then* $x^{k+1} \le \xi^{k+1}$*.*

We say that x *is the largest maximizer of* f*, denoted by* $\underset{\xi \in K}{\text{largmax}}\{f(\xi)\}$*, if it is a maximizer that satisfies* (ii) *and* (iii) *above with the inequalities reversed.*

The first question that one might ask is whether or not the sargmax is well defined for all functions in the Skorohod space. Before attempting to give an answer, we will use our notation to clarify the concept of a maximizer: a point $x \in K$ is a maximizer of $f \in \mathcal{D}_K$ if

$$\max_{\rho \in \mathcal{V}}\{f(x + 0_\rho)\} = \sup_{\xi \in K}\{f(\xi)\}.$$

We can now prove a result concerning the set of maximizers of a function in $\mathcal{D}_K$.

Lemma 4.2.7 *The set of maximizers of any function in* $\mathcal{D}_K$ *is compact.*

Proof: Let $f \in \mathcal{D}_K$. Since the set of maximizers of f is a subset of the compact rectangle K, it suffices to show that any convergent sequence of maximizers converges to a maximizer. Let $(x_n)_{n=1}^\infty$ be a sequence of maximizers with limit x. For each x_n we can find ξ_n with $|x_n - \xi_n| < \frac{1}{n}$ and such that $|f(\xi_n) - \max_{\rho \in \mathcal{V}}\{f(x_n + 0_\rho)\}| < 1/n$. Then we have that $\xi_n \to x$ and $|f(\xi_n) - \sup_{\xi \in K}\{f(\xi)\}| < 1/n \; \forall \; n \in \mathbb{N}$. Since K is the disjoint union of $\{\tilde{Q}(\rho, x)\}_{\rho \in \mathcal{V}}$, it follows that there is $\rho_* \in \mathcal{V}$ and a subsequence $(\xi_{n_k})_{k=1}^\infty$ such that $\xi_{n_k} \in \tilde{Q}(\rho_*, x) \; \forall \; k \in \mathbb{N}$. Therefore, the remark stated right after the definition of the Skorohod space implies that $f(\xi_{n_k}) \to f(x + 0_{\rho_*})$ and, consequently, $f(x+0_{\rho_*}) = \sup_{\xi \in K}\{f(\xi)\}$. $\square$

The previous lemma can be used to show that the sargmax functional is well defined on $\mathcal{D}_K$.

Lemma 4.2.8 *For each $f \in \mathcal{D}_K$ there is a unique element in $x \in K$ such that $x = \underset{\xi \in K}{\mathit{sargmax}}\{f(\xi)\}$.*

Proof: Let $f \in \mathcal{D}_K$. Since the set of maximizers of f is compact, if we can show that it is nonempty then the compactness will imply that there is a unique element $x \in K$ satisfying properties (i), (ii) and (iii) of Definition 4.2.4. Hence, it suffices to show that f has at least one maximizer. For this purpose, for each $n \in \mathbb{N}$ choose x_n such that $\sup_{\xi \in K}\{f(\xi)\} < f(x_n) + \frac{1}{n}$. Since K is compact, there is $x \in K$ and a subsequence $(x_{n_k})_{k=1}^{\infty}$ such that $x_{n_k} \to x$. Just as in the proof of the previous lemma, we can find $\rho_* \in \mathcal{V}$ and a further subsequence $(x_{n_{k_s}})_{s=1}^{\infty}$ such that $x_{n_{k_s}} \in \tilde{Q}(\rho_*, x)\ \forall\ s \in \mathbb{N}$. It follows that $f(x_{n_{k_s}}) \to f(x+0_{\rho_*})$ and hence $\sup_{\xi \in K}\{f(\xi)\} = f(x+0_{\rho_*})$. Therefore, the set of maximizers is nonempty and the sargmax is well defined. $\square$

We finish this section with a continuity theorem for the sargmax functional on continuous functions.

Lemma 4.2.9 *Let $W \in \mathcal{D}_K$ be a continuous function which has a unique maximizer $x^* \in K$. Then, the smallest argmax functional is continuous at W (with respect to d_K, $\widetilde{d}_K$ and the sup-norm metrics).*

Proof: Let $(W_n)_{n=1}^{\infty}$ be a sequence converging to W in the Skorohod topology. Let $\epsilon > 0$ be given and G be the open ball of radius ϵ around x^* and let $\delta := \left(W(x^*) - \sup_{x \in K \setminus G}\{W(x)\}\right)/2 > 0$. By Lemma 4.2.6 we have

$\|W_n - W\|_K < \delta$ for all large n (d_K, $\widetilde{d}_K$ and $\|\cdot\|_K$ generate the same local topology on W). Then

$$W(x^*) = 2\delta + \sup_{x \in K \setminus G} \{W(x)\} > \delta + \sup_{x \in K \setminus G} \{W_n(x)\}.$$

But $\|W_n - W\|_K < \delta$ also implies that $\sup_{x \in K}\{W_n(x)\} > W(x^*) - \delta$. The combination of these two facts shows that if $\|W_n - W\|_K < \delta$, then any maximizer of W_n must belong to G. Thus, $|\operatorname{sargmax}_{x \in K}\{W_n(x)\} - x^*| < \epsilon$ for n large enough. $\square$

4.3 A continuous mapping theorem for the sargmax functional on functions with jumps

Lemma 4.2.9 shows that the sargmax functional is continuous on continuous functions with unique maximizers. However, its raison d'être is to fix a unique maximizer on a function having multiple maximizers. Thus, a continuous mapping theorem on functions with jumps and possibly multiple maximizers is desired. We will show a version of the continuous mapping theorem on a suitable subset of our space $\widetilde{\mathcal{D}}_K$.

To state and prove our version of the continuous mapping theorem for the sargmax functional, we need to introduce some notation. We start with the space $\mathcal{D}_K^0$ consisting of all functions $\psi : K_1 \times K_2 \to \mathbb{R}$ which can be expressed as:

$$\psi(t, \xi) = V_0(\xi)\mathbf{1}_{a_{-1} \le t < a_1} + \sum_{k=1}^{\infty} V_k(\xi)\mathbf{1}_{a_k \le t < a_{k+1}} + \sum_{k=1}^{\infty} V_{-k}(\xi)\mathbf{1}_{a_{-k-1} \le t < a_{-k}} \tag{4.3}$$

where $(\ldots < a_{-k-1} < a_{-k} < \ldots < a_0 = 0 < \ldots < a_k < a_{k+1} < \ldots)_{k \in \mathbb{N}}$ is a sequence of jumps and $(V_k)_{k \in \mathbb{Z}}$ is a collection of continuous functions. Note that

$\mathcal{D}_K^0 \subset \widetilde{\mathcal{D}}_K$. Observe that the representation in (4.3) is not unique. However, knowledge of the function ψ and of the jumps $(a_k)_{k\in\mathbb{Z}}$ completely determines the continuous functions $(V_k)_{k\in\mathbb{Z}}$.

Our theorem will require not only Skorohod convergence of the elements of $\mathcal{D}_K^0$, but also convergence of their associated *pure jump functions*. To define properly these jump functions, we introduce the space $\mathcal{S}$ all piecewise constant, cádlág functions $\tilde{\psi} : \mathbb{R} \to \mathbb{R}$ such that $\tilde{\psi}(0) = 0$; $\tilde{\psi}$ has jumps of size 1; and $\tilde{\psi}(-t)$ and $\tilde{\psi}(t)$ are nondecreasing on $(0, \infty)$. For any closed interval $I \subset \mathbb{R}$ we introduce the space $\mathcal{S}_I := \{f|_I : f \in \mathcal{S}\}$. We endow the spaces $\mathcal{S}_I$ with the usual Skorohod topology d_I. Observe that the fact that all elements of $\mathcal{S}$ are cádlág and have jumps of size one implies that *any function in $\mathcal{S}_I$ has a finite number of jumps on I.*

We associate with every $\psi \in \mathcal{D}_K^0$, expressed as in (4.3), a pure jump function $\tilde{\psi} \in \mathcal{S}$ whose sequence of jumps is exactly the a_k's, i.e.,

$$\tilde{\psi}(t) = \sum_{k=1}^{\infty} \mathbf{1}_{a_k \le t} + \sum_{k=1}^{\infty} \mathbf{1}_{a_{-k} > t}. \tag{4.4}$$

We will show that Skorohod-convergence of functions in $\mathcal{D}_K^0$ and Skorohod convergence of their associated pure jump functions implies convergence of the corresponding sargmax and largmax functionals.

The following convergence result is a generalization of both, Lemma 3.1 of Lan et al. (2009) and Lemma A.3 in Seijo and Sen (2011a).

Theorem 4.3.1 *Assume that $d \ge 2$ and let $\left(\psi_n, \tilde{\psi}_n\right)_{n=1}^{\infty}$, $(\psi_0, \tilde{\psi}_0)$ be functions in $\mathcal{D}_K^0 \times \mathcal{S}_{K_1}$ such that ψ_n satisfies (4.3) for the sequence of jumps of $\tilde{\psi}_n$ for any $n \ge 0$. Assume that $(\psi_n, \tilde{\psi}_n) \to (\psi_0, \tilde{\psi}_0)$ in $\mathcal{D}_K^0 \times \mathcal{S}_{K_1}$ (with the product topology). Suppose, in addition, that ψ_0 can be expressed as (4.3) for the sequence of jumps $(\ldots < a_{-k-1} < a_{-k} < \ldots < a_0 = 0 < \ldots < a_k < a_{k+1} < \ldots)_{k\in\mathbb{N}}$ of $\tilde{\psi}_0$ and some continuous functions $(V_j)_{j\in\mathbb{Z}}$, each having a* *unique maximizer on K_2, with the property that for any finite subset $A \subset \mathbb{Z}$ there is only one $j \in A$ for which*

$$\max_{m\in A}\left\{\sup_{\xi\in K_2}\{V_m(\xi)\}\right\} = \sup_{\xi\in K_2}\{V_j(\xi)\}\,. \tag{4.5}$$

Finally, assume that ψ_0 has no jumps at the extreme points of K_1. Then,

(i) $\underset{x\in K}{\text{sargmax}}\{\psi_n(x)\} \to \underset{x\in K}{\text{sargmax}}\{\psi_0(x)\}$ *as* $n\to\infty$*;*

(ii) $\underset{x\in K}{\text{largmax}}\{\psi_n(x)\} \to \underset{x\in K}{\text{largmax}}\{\psi_0(x)\}$ *as* $n\to\infty$*.*

The result is also true when $d = 1$ under the same assumptions, but taking the sequence $(V_j)_{j\in\mathbb{Z}}$ to be a sequence of constants such that for any finite subset $A \subset \mathbb{Z}$ there is a unique $j \in A$ such that $\max_{m\in A}\{V_m\} = V_j$.

Proof: We focus on the case when $d > 1$ as the one-dimensional case is just Lemma 3.1 of Lan et al. (2009). Without loss of generality, assume that $K_1 = [-C, C]$ for some $C > 0$.

We can write ψ_n in the form (4.3) with $(\ldots < a_{n,-k-1} < a_{n,-k} < \ldots < a_{n,0} = 0 < \ldots < a_{n,k} < a_{n,k+1} < \ldots)_{k\in\mathbb{N}}$ being the sequence of jumps of ψ_n and $V_{n,j}$ being the continuous functions. Consequently, $\tilde{\psi}_n$, the pure jump function associated with ψ_n, can be expressed as (4.4) with jumps at $(a_{n,k})_{k\in\mathbb{Z}}$.

Let N_r and N_l be the number of jumps of $\tilde{\psi}_0$ in $[0, C]$ and $[-C, 0)$ respectively. Let $\epsilon > 0$ be sufficiently small such that all the points of the form $a_j \pm \epsilon$ are continuity points of ψ_0, for $-N_l \le j \le N_r$. Since convergence in the Skorohod topology of $\tilde{\psi}_n$ to $\tilde{\psi}_0$ implies point-wise convergence for continuity points of $\tilde{\psi}_0$ (see page 121 of Billingsley (1968)), and all of them are integer-valued functions, we see that $\tilde{\psi}_n(a_j - \epsilon) = j - 1$ and $\tilde{\psi}_n(a_j + \epsilon) = j$ for any $1 \le j \le N_r$, and $\tilde{\psi}_n(C) = N_r$ for all sufficiently large n. Thus, for all but finitely many n's we have that $\tilde{\psi}_n$ has exactly N_r jumps between 0 and C and that the location of the j-th jump to the right of 0 satisfies $|a_{n,j} - a_j| < \epsilon$.

Since $\epsilon > 0$ can be made arbitrarily small, we get that all the jumps $a_{n,j}$ converge to their corresponding a_j for all $1 \le j \le N_r$. The same happens to the left of zero: for all but finitely many n's, $\tilde{\psi}_n$ has exactly N_l jumps in $[-C, 0)$ and the sequences of jumps $(a_{n,-j})_{n=1}^{\infty}$, $1 \le j \le N_l$, converge to the corresponding jumps a_{-j}.

Let $V^* = \sup\{V_j(\xi) : \xi \in K_2, -N_l \le j \le N_r\}$. Our assumptions on the V_j's imply that this supremum is actually achieved at some unique vector $\xi^* \in K_2$ and that there is a unique "flat stretch" at which this supremum is attained (the last assertion follows form (4.5)).

Suppose, without loss of generality, that the maximum value is achieved in an interval of the form $[a_k, a_{k+1} \wedge C)$ for a unique $k \in \{1, \dots, N_r\}$. Now, write $b_0 = 0$; $b_j = \frac{a_j + C \wedge a_{j+1}}{2}$ for $1 \le j \le N_r$; and $b_j = \frac{a_j + (-C) \vee a_{j-1}}{2}$ for $-N_l \le j \le -1$. Note that the b_j's (for any value of $\xi \in K_2$) are continuity points of both ψ_0 and $\tilde{\psi}_0$.

Let $\kappa = \min_{-N_l \le j \le N_r+1}(C \wedge a_j - (-C) \vee a_{j-1})$ be the length of the shortest stretch. Take $0 < \eta, \delta < \kappa/4$. Considering the convergence of the jumps of ψ_n to those of ψ_0, there is $N \in \mathbb{N}$ such that for any $n \ge N$, the following two statements hold:

(a) Consider $\rho > 0$ such that if $|||\lambda|||_{K_1} < \rho$, then

$$\sup\{|s - \lambda(s)| : s \in [-C, C]\} < \delta.$$

The existence of such ρ follows from Lemma 4.2.5. By the convergence of ψ_n to ψ_0 in the Skorohod topology, there exists $\lambda_n \in \Lambda_{K_1}$ such that $|||\lambda_n|||_{K_1} < \rho$ and

$$\sup_{(t,\xi) \in K_1 \times K_2} \{|\psi_n(\lambda_n(t), \xi) - \psi_0(t, \xi)|\} < \eta.$$

(b) For any $1 \le j \le N_r$ (respectively, $j = 0$, $-N_l \le j \le -1$), b_j lies somewhere inside the interval $(a_{n,j} + \delta, C \wedge a_{n,j+1} - \delta)$ (respectively $(a_{n,-1} + \delta,$

$a_{n,1} - \delta), ((-C) \vee a_{n,j-1} + \delta, a_{n,j} - \delta))$. This follows from what was proven in the first two paragraphs of this proof.

From (a) we see that $|\lambda_n(b_j) - b_j| < \delta$ for all $-N_l \leq j \leq N_r$. But (b) and the size of δ in turn imply that b_j and $\lambda_n(b_j)$ belong to the same "flat stretch" of ψ_n and thus $\psi_n(\lambda_n(b_j), \xi) = \psi_n(b_j, \xi) = V_{n,j}(\xi)$ for all $\xi \in K_2$ and all $-N_l \leq j \leq N_r$. Considering again (b) and the second inequality in (a), we conclude that $\|V_{n,j} - V_j\|_{K_2} < \eta$ for all $-N_l \leq j \leq N_r$ and all $n \geq N$. Hence, all the sequences $(V_{n,j})_{n=1}^{\infty}$ converge uniformly in K_2 to their corresponding V_j. Consequently:

$$\begin{aligned}
\max_{\substack{-N_l \leq j \leq N_r \\ j \neq k}} \left\{ \sup_{\xi \in K_2} V_{n,j}(\xi) \right\} &\longrightarrow \max_{\substack{-N_l \leq j \leq N_r \\ j \neq k}} \left\{ \sup_{\xi \in K_2} V_j(\xi) \right\}, \\
\max_{\xi \in K_2} \{V_{n,k}(\xi)\} &\longrightarrow \max_{\xi \in K_2} \{V_k(\xi)\} = V_k(\xi^*), \\
\underset{\xi \in K_2}{\operatorname{argmax}} \{V_{n,k}(h_1, h_2)\} &\longrightarrow \underset{\xi \in K_2}{\operatorname{argmax}} \{V_k(\xi)\} = \xi^*, \\
\overline{\lim_{n \to \infty}} \max_{\substack{-N_l \leq j \leq N_r \\ j \neq k}} \left\{ \sup_{\xi \in K_2} V_{n,j}(\xi) \right\} &< \underline{\lim_{n \to \infty}} \max_{\xi \in K_2} \{V_{n,k}(\xi)\}.
\end{aligned}$$

The above, together with (4.5) and the fact that $a_{n,k} \to a_k$ and $a_{n,k+1} \to a_{k+1}$, imply that

$$\underset{x \in K}{\operatorname{sargmax}} \{\psi_n(x)\} \to (\xi^*, a_k) = \underset{x \in K}{\operatorname{sargmax}} \{\psi_0(x)\}$$

$$\underset{x \in K}{\operatorname{largmax}} \{\psi_n(x)\} \to (\xi^*, a_{k+1}) = \underset{x \in K}{\operatorname{largmax}} \{\psi_0(x)\}$$

as $n \to \infty$. $\square$

We now present a version of the previous result but for random elements in $\mathcal{D}_K^0$. To prove it, we will use Lemma 4.2 in Prakasa Rao (1969) (which we present here to ease the exposition). In the remaining of the paper we will use the symbol $\rightsquigarrow$ to represent weak convergence.

Lemma 4.3.1 (Prakasa Rao (1969)) *Consider the random vectors* $\{W_{n\epsilon}, W_n, W_\epsilon\}_{\epsilon\geq 0}^{n\in\mathbb{N}}$ *and* W*. Suppose that the following conditions hold:*

(i) $\lim_{\epsilon\to 0} \overline{\lim_{n\to\infty}}\, \mathbf{P}\,(W_{n\epsilon} \neq W_n) = 0$,

(ii) $\lim_{\epsilon\to 0} \mathbf{P}\,(W_\epsilon \neq W) = 0$,

(iii) $W_{n\epsilon} \rightsquigarrow W_\epsilon$ *(as* $n \to \infty$*) for every* $\epsilon > 0$.

Then, $W_n \rightsquigarrow W$.

In the next theorem we will be taking the sargmax and largmax functionals over rectangles that may not be compact. When this happens, we say that these functionals are *well defined* if there is an element in the corresponding rectangle satisfying conditions $(i) - (iii)$ defining the smallest and largest argmax functionals (see Definition 4.2.4). If we are given a rectangle $\Theta \subset \mathbb{R}^d$ which can be written as the Cartesian product of possibly unbounded closed intervals, we will denote by $\mathcal{D}_\Theta$ the collection of functions $f : \Theta \to \mathbb{R}$ whose restrictions to all compact rectangles $K \subset \Theta$ belong to $\mathcal{D}_K$.

Theorem 4.3.2 *Assume that* $K = K_1 \times K_2$ *is a closed rectangle in* $\mathbb{R}^d$ *and that* $0 \in K_1^\circ \subset \mathbb{R}$*. Let* $(\Omega, \mathcal{F}, \mathbf{P})$ *be a probability space and let* $(\Psi_n, \Gamma_n)_{n=1}^\infty$, (Ψ_0, Γ_0) *be random elements taking values in* $\mathcal{D}_K^0 \times \mathcal{S}_{K_1}$ *such that* Ψ_n *satisfies (4.3) for the sequence of jumps of* Γ_n *for any* $n \geq 0$*, almost surely. Moreover, suppose that, with probability one, we have that:* Ψ_0 *satisfies (4.5);* Γ_0 *has no fixed time of discontinuity; the sargmax and largmax functionals over* K *are finite for* Ψ_0 *(this assumption is essential as* K *is not necessarily compact). If the following hold:*

(i) For every compact subinterval $B_1 \subset K_1$ *and compact sub-rectangle* $B := B_1 \times B_2 \subset K$ *we have* $(\Psi_n, \Gamma_n) \rightsquigarrow (\Psi_0, \Gamma_0)$ *on* $\mathcal{D}_B \times \mathcal{D}_{B_1}$*;*

(ii) $\left(\underset{\theta\in K}{\mathit{sargmax}}\{\Psi_n(\theta)\}, \underset{\theta\in K}{\mathit{largmax}}\{\Psi_n(\theta)\}\right) = O_{\mathbf{P}}(1)$;

then we also have

$$\left(\underset{\theta\in K}{\mathit{sargmax}}\{\Psi_n(\theta)\}, \underset{\theta\in K}{\mathit{largmax}}\{\Psi_n(\theta)\}\right) \rightsquigarrow \left(\underset{\theta\in K}{\mathit{sargmax}}\{\Psi_0(\theta)\}, \underset{\theta\in K}{\mathit{largmax}}\{\Psi_0(\theta)\}\right).$$

Proof: Consider $C > 0$ and let

$$\begin{aligned} \phi_n &:= \left(\underset{\theta\in K}{\operatorname{sargmax}}\{\Psi_n(\theta)\}, \underset{\theta\in K}{\operatorname{largmax}}\{\Psi_n(\theta)\}\right) \\ \phi_{n,C} &:= \left(\underset{\theta\in[-C,C]^d\cap K}{\operatorname{sargmax}}\{\Psi_n(\theta)\}, \underset{\theta\in[-C,C]^d\cap K}{\operatorname{largmax}}\{\Psi_n(\theta)\}\right), \end{aligned}$$

for all $n \geq 0$. To prove the result, we will apply Theorem 4.3.1 and Lemma 4.3.1. Using the notation of the latter, set $\epsilon = \frac{1}{C}$, $W_{n\epsilon} = \phi_{n,C}$ for $n \geq 1$, $W_\epsilon = \phi_{0,C}$, $W_n = \phi_n$ for $n \geq 1$ and $W = \phi_0$. From (ii) we see that $\lim_{\epsilon\to 0} \overline{\lim_{n\to\infty}} \mathbf{P}\left(W_{n\epsilon} \neq W_n\right) = 0$. Our assumptions on Ψ_0 and Γ_0 imply that $\lim_{\epsilon\to 0} \mathbf{P}\left(W_\epsilon \neq W\right) = 0$. Finally, Theorem 4.3.1 and an application of Skorohod's Representation Theorem (see either Theorem 1.8, page 102 in Ethier and Kurtz (2005) or Theorems 1.10.3 and 1.10.4, pages 58 and 59 in van der Vaart and Wellner (1996)) show that $W_{n\epsilon} \rightsquigarrow W_\epsilon$ and hence, from Lemma 4.3.1, we conclude that $\phi_n \rightsquigarrow \phi_0$. $\square$

4.4 On the necessity of the convergence of the associated pure jump processes

Condition (i) in Theorem 4.3.2 involves the joint convergence of the processes whose maximizers are being considered and their associated pure jump processes. One may ask whether or not this condition is actually necessary for the weak convergence of the corresponding smallest maximizers. A simple counterexample shows that such a condition is indeed essential to guarantee the desired weak convergence under the assumptions of Theorem 4.3.2.

Let Ψ be a two-sided, right-continuous Poisson process and $T_{\pm 1} := \pm \inf\{t > 0 : \Psi(\pm t) > 0\}$. Consider the following $\mathcal{D}_{\mathbb{R}}$-valued random elements: $\Psi_0 := -\Psi$ and $\Psi_n = \Psi_0 + \frac{1}{n}\mathbf{1}_{\left[\frac{1}{2}T_{-1}, \frac{1}{2}T_1\right)}$. Then, $\Psi_n \rightsquigarrow \Psi$ in $\mathcal{D}_I$ for every compact interval I (in fact, the weak convergence holds in $\mathcal{D}_{\mathbb{R}}$ with the corresponding Skorohod topology). However,

$$\left(\underset{\mathbb{R}}{\operatorname{sargmax}}\{\Psi_n\}, \underset{\mathbb{R}}{\operatorname{largmax}}\{\Psi_n\}\right) = \frac{1}{2}\left(\underset{\mathbb{R}}{\operatorname{sargmax}}\{\Psi_0\}, \underset{\mathbb{R}}{\operatorname{largmax}}\{\Psi_0\}\right),$$

for all $n \in \mathbb{N}$. It is easily seen that all the conditions of Theorem 3.2 hold, with the exception of (i). Hence, the weak convergence of the processes Ψ_n alone is not enough to guarantee weak convergence of the corresponding maximizers.

4.5 Applications

4.5.1 Stochastic design change-point regression

We start by analyzing the example of the least squares change-point estimator given by (4.2) in the Introduction. Assume that we are given an i.i.d. sequence of random vectors $\{X_n = (Y_n, Z_n)\}_{n=1}^{\infty}$ defined on a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ having a common distribution $\mathbb{P}$ satisfying (4.1) for some parameter $\theta_0 := (\zeta_0, \alpha_0, \beta_0) \in \Theta := [c_1, c_2] \times \mathbb{R}^2$. Suppose that Z has a uniformly bounded, strictly positive density f (with respect to the Lebesgue measure) on $[c_1, c_2]$ such that $\inf_{|z-\zeta_0|\leq\eta} f(z) > \kappa > 0$ for some $\eta > 0$ and that $\mathbb{P}(Z < c_1) \wedge \mathbb{P}(Z > c_2) > 0$. For $\theta = (\zeta, \alpha, \beta) \in \Theta$, $x = (y, z) \in \mathbb{R}^2$ write

$$m_\theta(x) := -\left(y - \alpha\mathbf{1}_{z\leq\zeta} - \beta\mathbf{1}_{z>\zeta}\right)^2,$$

and $\mathbb{P}_n$ for the empirical measure defined by $X_1, \ldots, X_n$. Note that $M_n(\theta) := -\mathbb{P}_n[m_\theta]$ and recall the definition of $\hat{\theta}_n$.

The asymptotic properties of this estimator are well-known and have been deduced by several authors. They are available, for instance, in Kosorok (2008b) or Seijo and Sen (2011a). It follows from Proposition 5.3.2 that $\sqrt{n}(\hat{\alpha}_n - \alpha_0) = O_{\mathbf{P}}(1)$, $\sqrt{n}(\hat{\beta}_n - \beta_0) = O_{\mathbf{P}}(1)$ and $n(\hat{\zeta}_n - \zeta_0) = O_{\mathbf{P}}(1)$.

For $h = (h_1, h_2, h_3) \in \mathbb{R}^3$, let $\vartheta_{n,h} := \theta_0 + \left(\frac{h_1}{n}, \frac{h_2}{\sqrt{n}}, \frac{h_3}{\sqrt{n}}\right)$ and

$$\hat{E}_n(h) := n\mathbb{P}_n\left[m_{\vartheta_{n,h}} - m_{\theta_0}\right].$$

A consequence of this rate of convergence result is that with probability tending to one, we have

$$\hat{h}_n := \operatorname*{sargmax}_{h\in\mathbb{R}^3} \hat{E}_n(h) = \left(n(\hat{\zeta}_n - \zeta_0), \sqrt{n}(\hat{\alpha}_n - \alpha_0), \sqrt{n}(\hat{\beta}_n - \beta_0)\right).$$

Write $\hat{J}_n$ for the pure jump process associated with $\hat{E}_n$. It is shown in Lemma 5.3.3 that

(a) $(\hat{E}_n, \hat{J}_n) \rightsquigarrow (E^*, J^*)$ in $\mathcal{D}_K \times \mathcal{S}_I$,

on every compact rectangle $K = I \times A \times B \subset \mathbb{R}^3$ for some process $E^* \in \mathcal{D}_{\mathbb{R}^3}$ with an associated pure jump process J^*. Then, an application of Theorem 4.3.2 shows that $\hat{h}_n = \left(n(\hat{\zeta}_n - \zeta_0), \sqrt{n}(\hat{\alpha}_n - \alpha_0), \sqrt{n}(\hat{\beta}_n - \beta_0)\right) \rightsquigarrow \operatorname*{sargmax}_{h\in\mathbb{R}^3}\{E^*(h)\}$, see Corollary 5.3.1.

We would like to point out that the derivation of the asymptotic distribution of this estimator can also be found in Kosorok (2008b). The arguments there can be modified to obtain the result from an application of Theorem 4.3.2.

4.5.2 Estimation in a Cox regression model with a change-point in time

Define $\Theta := (0,1) \times \mathbb{R}^{p+2q}$ for given $p, q \in \mathbb{N}$. For $\theta = (\tau, \xi) = (\tau, \alpha, \beta, \gamma) \in \Theta = (0,1) \times \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^q$ consider a survival time T^0, a censoring time C

and covariate cáglád (left-continuous with right-hand side limits) $\mathbb{R}^{p+q}$-valued process $Z = (Z_1, Z_2)$ where the sample paths of Z_1 and Z_2 live in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. Assume that C and Z have laws G and H, respectively. Note that G is a distribution on the nonnegative real line and H a probability measure on the space of left continuous processes with right-hand side limits. In our Cox model with a change-point in time we make the additional assumption that, conditionally on Z, the hazard function of the survival time is given by:

$$\begin{aligned} \lambda(t|Z) &:= \lim_{\Delta t \downarrow 0} \frac{\mathbf{P}\left(t \leq T^0 < t + \Delta t | T^0 \geq t;\ Z(s),\ 0 \leq s \leq t\right)}{\Delta t} \\ &= \lambda(t) e^{\alpha \cdot Z_1(t) + (\beta + \gamma \mathbf{1}_{t > \tau}) \cdot Z_2(t)} \end{aligned}$$

where λ is the *baseline hazard function* and $\cdot$ denotes the standard inner product on Euclidian spaces. We write $\mathbb{P}_{\theta,\lambda,G,H}$ for the law of (T^0, C, Z). We would like to point out that we assume that G and the finite dimensional distributions of Z are all *continuous*.

Suppose that there is a random sample

$$(T_1^0, C_1, Z_{1,1}, Z_{2,1}), \ldots, (T_n^0, C_n, Z_{1,n}, Z_{2,n}) \overset{i.i.d.}{\sim} \mathbb{P}_{\theta_0, \lambda_0, G_0, H_0}$$

from which we are only able to observe $Z_{1,j},\ Z_{2,j},\ \Delta_j := \mathbf{1}_{T_j^0 \leq C_j}$ and $T_j := T_j^0 \wedge C_j$ for $j = 1, \ldots, n$. The goal is to estimate the change-point $\tau_0 \in (0,1)$ given these observations.

A standard method of estimation in this setting is via Cox's partial likelihood, in which case the likelihood and log-likelihood functions are given by

$$\begin{aligned} L_n(\tau, \alpha, \beta, \gamma) &:= \prod_{\substack{1 \leq k \leq n \\ T_k^0 \leq C_k}} \frac{e^{\alpha \cdot Z_{1,k}(T_k^0) + (\beta + \gamma \mathbf{1}_{T_k^0 > \tau}) \cdot Z_{2,k}(T_k^0)}}{\sum_{\{1 \leq j \leq n:\ T_k^0 \leq T_j^0 \wedge C_j\}} e^{\alpha \cdot Z_{1,j}(T_k^0) + (\beta + \gamma \mathbf{1}_{T_k^0 > \tau}) \cdot Z_{2,j}(T_k^0)}}, \\ l_n(\theta) &:= \log\left(L_n(\tau, \xi)\right) = \log\left(L_n(\tau, \alpha, \beta, \gamma)\right). \end{aligned}$$

In this case, the maximum partial likelihood estimator of the change-point and the covariate multipliers is given by

$$\hat{\theta}_n = (\hat{\tau}_n, \hat{\xi}_n) = (\hat{\tau}_n, \hat{\alpha}_n, \hat{\beta}_n, \hat{\gamma}_n) := \underset{\theta\in\Theta}{\text{sargmax}}\{l_n(\theta)\}.$$

Pons (2002) derived the asymptotics for this estimator. For $u = (u^1, u^2, \ldots, u^{1+p+2q}) = (u^1, v) \in \mathbb{R}^{1+p+2q}$ define $\theta_{n,u} = \left(\tau_0 + \frac{u^1}{n}, \xi_0 + \frac{v}{\sqrt{n}}\right)$. Then, under some regularity conditions, Theorem 2 in Pons (2002) shows that

$$\left(n(\hat{\tau}_n - \tau_0), \sqrt{n}(\hat{\xi}_n - \xi_0)\right) = \underset{u\in\mathbb{R}^{1+p+2q}:\ \theta_{n,u}\in\Theta}{\text{sargmax}}\{l_n(\theta_{n,u}) - l_n(\theta_0)\} = O_{\mathbf{P}}(1).$$

It can also be inferred from Proposition 3 and Theorem 3 of the same paper that $\Psi_n := l_n(\theta_{n,u}) - l_n(\theta_0) \rightsquigarrow \Psi$ on $\mathcal{D}_K$ for every compact rectangle $K \subset \mathbb{R}^{1+p+2q}$, where Ψ is a stochastic process of the form

$$\Psi(u^1, v) = Q(u^1) + v \cdot \tilde{W} - \frac{1}{2} v\tilde{I} \cdot v, \tag{4.6}$$

with Q being a two-sided, compound Poisson process, $\tilde{W}$ a Gaussian random variable independent of Q and $\tilde{I}$ some positive definite matrix on $\mathbb{R}^{(p+2q)\times(p+2q)}$. For a detailed description of Q, $\tilde{W}$ and $\tilde{I}$ we refer the reader to Section 4 of Pons (2002).

If one defines Γ_n and Γ to be the pure jump processes associated with Ψ_n and Ψ, respectively, it can be shown, using similar techniques as in the proof of Theorem 3 of Pons (2002), that $(\Psi_n, \Gamma_n) \rightsquigarrow (\Psi, \Gamma)$ on $\mathcal{D}_B \times \mathcal{D}_{B_1}$ for every compact subinterval $B_1 \subset \mathbb{R}$ and compact rectangle $B := B_1 \times B_2 \subset \mathbb{R}^{1+p+2q}$. Hence, Theorem 4.3.2 can be applied in this situation to conclude that

$$\left(n(\hat{\tau}_n - \tau_0), \sqrt{n}(\hat{\xi}_n - \xi_0)\right) \rightsquigarrow \underset{u\in\mathbb{R}^{1+p+2q}}{\text{sargmax}}\{\Psi(u)\}.$$

It must be noted that the proof of Theorem 4 in Pons (2002) makes no mention of the pure jump processes Γ_n and Γ. On the second sentence of this proof, the author claims that the asymptotic distribution follows just from the weak convergence of the processes Ψ_n. As we saw in Section 4.4 this fact alone is not enough to conclude the weak convergence of the smallest maximizers. Thus, the argument given in this section completes the mentioned proof in Pons (2002).

4.5.3 Estimating a change-point in a Cox regression model according to a threshold in a covariate

We will now discuss another application from survival analysis. Consider again a Cox regression model but now with a covariate process of the form $Z = (Z_1, Z_2, Z_3)$ where Z_1 and Z_2 are as in Section 4.5.2 and Z_3 is a continuous random variable in $\mathbb{R}$. We will denote the survival and censoring times as in Section 4.5.2. We are now concerned with a hazard function of the form

$$\lambda(t|Z) = \lambda(t) e^{\alpha\cdot Z_1(t)+\beta\cdot Z_2(t)\mathbf{1}_{Z_3\leq\zeta}+\gamma\cdot Z_2(t)\mathbf{1}_{Z_3>\zeta}},$$

for $\alpha \in \mathbb{R}^q$, $\beta, \gamma \in \mathbb{R}^q$ and some $\zeta \in I$ where I is a closed interval entirely contained in the interior of the support of Z_3. We now consider the parameter space $\Theta := I \times \mathbb{R}^{p+2q}$ and we write $\theta = (\zeta, \xi) := (\zeta, \alpha, \beta, \gamma) \in \Theta$. The partial likelihood and log-likelihood functions are now given by

$$\begin{aligned} L_n(\zeta,\alpha,\beta,\gamma) &:= \prod_{\substack{1\leq k\leq n\\ T_k^0\leq C_k}} \frac{e^{\alpha\cdot Z_{1,k}(T_k^0)+\beta\cdot Z_{2,k}(T_k^0)\mathbf{1}_{Z_{3,k}\leq\zeta}+\gamma\cdot Z_{2,k}(T_k^0)\mathbf{1}_{Z_{3,k}>\zeta}}}{\sum_{\{1\leq j\leq n:\ T_k^0\leq T_j^0\wedge C_j\}} e^{\alpha\cdot Z_{1,j}(T_k^0)+\beta\cdot Z_{2,j}(T_k^0)\mathbf{1}_{Z_{3,j}\leq\zeta}+\gamma\cdot Z_{2,j}(T_k^0)\mathbf{1}_{Z_{3,j}>\zeta}}}, \\ l_n(\theta) &:= \log\left(L_n(\zeta,\xi)\right) = \log\left(L_n(\zeta,\alpha,\beta,\gamma)\right). \end{aligned}$$

As before, we assume that the observations come from a model with some specific value $\theta_0 \in \Theta$. Following the notation of Section 4.5.2, for $u = (u^1, u^2, \ldots, u^{1+p+2q}) = (u^1, v) \in \mathbb{R}^{1+p+2q}$ define $\theta_{n,u} = \left(\zeta_0 + \frac{u^1}{n}, \xi_0 + \frac{v}{\sqrt{n}}\right)$. Then, under some regularity conditions, Theorem 2 in Pons (2003) shows that

$$\left(n(\hat{\zeta}_n - \zeta_0), \sqrt{n}(\hat{\xi}_n - \xi_0)\right) = \underset{u \in \mathbb{R}^{1+p+2q}:\ \theta_{n,u} \in \Theta}{\operatorname{sargmax}} \{l_n(\theta_{n,u}) - l_n(\theta_0)\} = O_{\mathbf{P}}(1).$$

Lemma 5 and Theorem 3 in Pons (2003) show that $\Psi_n := l_n(\theta_{n,u}) - l_n(\theta_0) \rightsquigarrow \Psi$ on $\mathcal{D}_K$ for every compact rectangle $K \subset \mathbb{R}^{1+p+2q}$, where Ψ is another stochastic process of the form (4.6) but with different two-sided, compound Poisson process Q, Gaussian random variable $\tilde{W}$ and positive definite matrix $\tilde{I}$. The details can be found in Section 4 of Pons (2003).

Letting Γ_n and Γ to be the pure jump processes associated with Ψ_n and Ψ, respectively, it can be shown that $(\Psi_n, \Gamma_n) \rightsquigarrow (\Psi, \Gamma)$ on $\mathcal{D}_B \times \mathcal{D}_{B_1}$ for every compact subinterval $B_1 \subset \mathbb{R}$ and compact rectangle $B := B_1 \times B_2 \subset \mathbb{R}^{1+p+2q}$. Hence, another application of Theorem 4.3.2 shows that

$$\left(n(\hat{\tau}_n - \tau_0), \sqrt{n}(\hat{\xi}_n - \xi_0)\right) \rightsquigarrow \underset{u \in \mathbb{R}^{1+p+2q}}{\operatorname{sargmax}} \{\Psi(u)\}.$$

As in Pons (2002), the argument to derive the asymptotic distribution given in the proof of Theorem 5 lacks a proper discussion of the convergence of the associated pure jump processes. Therefore, the analysis just given can be seen as a complement to the proof of Theorem 5 in Pons (2003).

More general models involving right censoring for survival times and a change-point based on a threshold in a covariate can be found in Kosorok and Song (2007). There, the change-point estimator also achieves a n^{-1} rate of convergence. The asymptotic distribution of this estimator also corresponds to the smallest maximizer of a two-sided, compound Poisson process and can be deduced from an application of Theorem 4.3.2. We would like to point out that the above authors omit a discussion about the associated pure jump processes. They claim the desired stochastic convergence follows from an application of Theorem 3.2.2 in van der Vaart and Wellner (1996) (see the last paragraph of the proof of Theorem 5 in page 985 of Kosorok and Song (2007)), but this theorem cannot be applied as the maximizer of a compound Poisson process is not unique. Thus, a proper application of Theorem 4.3.2 would complete the argument in Kosorok and Song (2007).

Chapter 5

Change-point regression and the bootstrap

5.1 Introduction

Change-point models may arise when a stochastic system is subject to sudden external influences and are encountered in almost every field of science. In the simplest form the model considers a random vector $X = (Y, Z)$ satisfying the following relation:

$$Y = \alpha_0 \mathbf{1}_{Z \le \zeta_0} + \beta_0 \mathbf{1}_{Z > \zeta_0} + \epsilon, \tag{5.1}$$

where Z is a continuous random variable, $\alpha_0 \neq \beta_0 \in \mathbb{R}$, $\zeta_0 \in [a, b] \subset \mathbb{R}$ and ϵ is a continuous random variable, independent of Z with zero expectation and finite variance $\sigma^2 > 0$. The parameter of interest is ζ_0, the change-point.

Despite its simplicity, model (5.1) captures the inherent "non-standard" nature of the problem: The least squares estimator of the change-point ζ_0 converges at a rate of n^{-1} to a minimizer of a two-sided, compound Poisson process that depends crucially on the entire error distribution, the marginal density of Z, among other nuisance parameters; see Pons (2003), Kosorok (2008b) (Section 14.5.1, pages 271–277) or Koul et al. (2003). Therefore, it is not practical to use this limiting distribution to build a confidence interval (CI) for ζ_0. Bootstrap methods bypass the estimation of nuisance parameters and are generally reliable in $\sqrt{n}$-convergence problems. In this chapter we investigate the performance (both theoretically and through simulation) of different bootstrap schemes to build CIs for ζ_0. We hope our analysis will help illustrate the issues that arise when the bootstrap is applied in such non-standard problems.

The problem of estimating a jump-discontinuity (change-point) in an otherwise smooth curve has been under study for at least the last forty years. More recently, it has been extensively studied in the nonparametric regression and survival analysis literature; see for instance Gijbels et al. (1999), Dempfle and Stute (2002), Pons (2003), Kosorok and Song (2007), Lan et al. (2009) and the references therein. Bootstrap techniques have also been applied in many instances in change point models. Dümbgen (1991) proposed asymptotically valid confidence regions for the change-point by inverting bootstrap tests in a one-sample problem. Hŭsková and Kirch (2008) considered bootstrap CIs for the change-point of the mean in a time series context. Kosorok and Song (2007) use a form of parametric bootstrap to estimate the distribution of the estimated change-point in a stochastic design regression model that arises in survival analysis. Gijbels et al. (2004), in a slightly different setting, suggested a bootstrap procedure for model (5.1), but did not give a complete proof of its validity.

Our work goes beyond those cited above as follows: We present strong theoretical and empirical evidence to suggest the *inconsistency* of the two most natural bootstrap procedures in a regression setup – the usual nonparametric bootstrap (i.e., sampling from the empirical cumulative distribution function (ECDF) of (Y, Z), often also called as bootstrapping "pairs") and the "residual" bootstrap. The bootstrap estimators built by both of these methods are the smallest maximizers of certain stochastic processes. We show that these processes do not have any weak limit in probability. This fact strongly suggests not only inconsistency but also the absence of *any* weak limit for the bootstrap estimators. In addition, we prove that independent sampling from a smooth approximation to the marginal of Z and the centered ECDF of the residuals, and the m out of n bootstrap from the ECDF of (Y, Z) yield asymptotically valid CIs for ζ_0. The finite sample performance of the different bootstrap methods shows the superiority of the proposed smoothed bootstrap procedure. We also develop a series of convergence results which generalize those obtained in Kosorok (2008b) to triangular arrays of random vectors and can be used to validate the consistency of *any* bootstrap scheme in this setup.

Although we develop our results in the setting of (5.1), our conclusions have broader implications (as discussed in Section 5.7). They extend immediately to regression functions with parametrically specified models on either side of the change-point. The *smoothed bootstrap* procedure can also be modified to work in more general nonparametric settings. Gijbels et al. (1999) consider jump-point estimation in the more general setup of non-parametric regression and develop two-stage procedures to build CI for the change-point. In the second stage of their procedure, they localize to a neighborhood of the change-point and reduce the problem to exactly that of (5.1). Lan et al. (2009) consider a two-stage adaptive sampling procedure to estimate the jump discontinuity. The second stage of their method relies on an approximate CI for the change-point, and the bootstrap methods developed here can be immediately used in their context.

The chapter is organized in the following manner: In Section 5.2 we describe the problem in greater detail, introduce the bootstrap schemes and describe the appropriate notion of consistency. In Section 5.3, we prove a series of convergence results that generalize those obtained in Kosorok (2008b). These results will constitute the general framework under which the bootstrap schemes will be analyzed. In Section 5.4 we study the inconsistency of the standard bootstrap methods, including the ECDF and residual bootstraps. In Section 5.5 we propose two bootstrap procedures and show their consistency. We compare the finite sample performance of the different bootstrap methods through a simulation study in Section 5.6. Finally, in Section 5.7 we discuss the consequences of our analysis in more general change-point regression models. To ease the exposition, we have relegated many of the proof and some auxiliary results to Section 5.9.

5.2 The problem and the bootstrap schemes

Assume that we are given an i.i.d. sequence of random vectors $\{X_n = (Y_n, Z_n)\}_{n=1}^{\infty}$ defined on a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ having a common distribution $\mathbb{P}$ satisfying (5.1) for some parameter $\theta_0 := (\alpha_0, \beta_0, \zeta_0) \in \Theta := \mathbb{R}^2 \times [a, b]$. This is a semi-parametric model with an Euclidean parameter θ_0 and two infinite-dimensional parameters – the distributions of Z and ϵ. We are interested in estimating ζ_0, the change-point. For technical reasons, we will also assume that $\mathbb{P}(|\epsilon|^3) < \infty$. Here, and in the remaining of the paper, we take the convention that for any probability distribution μ, we will denote the expectation operator by $\mu(\cdot)$. In addition, we suppose that Z has a uniformly bounded, strictly positive density f (with respect to the Lebesgue measure) on $[a, b]$ such that $\inf_{|z-\zeta_0|\leq\eta} f(z) > \kappa > 0$ for some $\eta > 0$ and that $\mathbb{P}(Z < a) \wedge \mathbb{P}(Z > b) > 0$.

For $\theta = (\zeta, \alpha, \beta) \in \Theta$, $x = (y, z) \in \mathbb{R}^2$ write

$$m_\theta(x) := -\left(y - \alpha \mathbf{1}_{z\le\zeta} - \beta \mathbf{1}_{z>\zeta}\right)^2, \tag{5.2}$$

$\mathbb{P}_n$ for the empirical measure defined by $X_1, \ldots, X_n$,

$$M_n(\theta) := \mathbb{P}_n(m_\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \alpha \mathbf{1}_{Z_i\le\zeta} + \beta \mathbf{1}_{Z_i>\zeta}\right)^2, \tag{5.3}$$

and $M(\theta) := \mathbb{P}(m_\theta)$. The function M_n is strictly concave in its first two coordinates but càdlàg (right continuous with left limits) in the third; in fact, piecewise constant and with n jumps (w.p. 1). Thus, M_n has unique maximizing values of α and β, but an entire interval of maximizers for ζ. For this reason, we define the *least squares estimator* of θ_0 to be the maximizer of M_n over Θ with the smallest ζ, and denote it by

$$\hat{\theta}_n := (\hat{\zeta}_n, \hat{\alpha}_n, \hat{\beta}_n) = \underset{\theta\in\Theta}{\operatorname{sargmax}}\left\{M_n(\theta)\right\},$$

where sargmax stands for the *smallest argmax functional*, as defined in 4.2.4.

The asymptotic properties of this least squares estimator are well known. It is shown in Kosorok (2008b), pages 271–277, that $\sqrt{n}(\hat{\alpha}_n - \alpha_0) = O_{\mathbf{P}}(1)$, $\sqrt{n}(\hat{\beta}_n - \beta_0) = O_{\mathbf{P}}(1)$ and $n(\hat{\zeta}_n - \zeta_0) = O_{\mathbf{P}}(1)$. The asymptotic distribution of $n(\hat{\zeta}_n - \zeta_0)$ is that of the smallest argmax of a two-sided compound Poisson process. However, the limiting process depends on the distribution of ϵ and the value of the density of Z at ζ_0. Thus, there is no straightforward way to build CIs for ζ_0 using this limiting distribution. In this connection we investigate the performance of bootstrap procedures for constructing CIs for ζ_0.

5.2.1 Bootstrap

We start with a brief review of the bootstrap. Given a sample $\mathbf{W}_n = \{W_1, W_2, \ldots, W_n\} \overset{\text{iid}}{\sim} L$ from an unknown distribution L, suppose that the distribution function H_n of some random variable $R_n \equiv R_n(\mathbf{W}_n, L)$ is of interest; R_n is usually called a *root* and it can in general be any measurable function of the data and the distribution L. The bootstrap method can be broken into three simple steps:

(i) Construct an estimator $\hat{L}_n$ of L from $\mathbf{W}_n$.

(ii) Generate $\mathbf{W}_n^* = \{W_1^*, \ldots, W_{m_n}^*\} \overset{\text{iid}}{\sim} \hat{L}_n$ given $\mathbf{W}_n$.

(iii) Estimate H_n by $\hat{H}_n$, the conditional CDF of $R_n(\mathbf{W}_n^*, \hat{L}_n)$ given $\mathbf{W}_n$.

Let d denote the Prokhorov metric or any other metric metrizing weak convergence of probability measures. We say that $\hat{H}_n$ is *weakly consistent* if $d(H_n, \hat{H}_n) \overset{P}{\to} 0$; if H_n has a weak limit H, this is equivalent to $\hat{H}_n$ converging weakly to H in probability. Similarly, $\hat{H}_n$ is *strongly consistent* if $d(H_n, \hat{H}_n) \overset{a.s.}{\to} 0$.

The choice of $\hat{L}_n$ mostly considered in the literature is the ECDF. Intuitively, an $\hat{L}_n$ that mimics the essential properties (e.g., smoothness) of the underlying distribution L can be expected to perform well. Despite being a good estimator in most situations, the ECDF can fail to capture some properties of L that may be crucial for the problem under consideration. This is especially true in nonstandard problems. In Section 5.4 we illustrate this phenomenon (the inconsistency of the ECDF bootstrap) when $n(\hat{\zeta}_n - \zeta_0)$ is the random variable (root) of interest.

We denote by $\mathfrak{X} = \sigma\left((X_n)_{n=1}^{\infty}\right)$ the σ-algebra generated by the sequence $(X_n)_{n=1}^{\infty}$ and write $\mathbf{P}_{\mathfrak{X}}(\cdot) = \mathbf{P}(\cdot\,|\mathfrak{X})$ and $\mathbf{E}_{\mathfrak{X}}(\cdot) = \mathbf{E}(\cdot\,|\mathfrak{X})$. We approximate the CDF of $\Delta_n = n(\hat{\zeta}_n - \zeta_0)$ by $\mathbf{P}_{\mathfrak{X}}(\Delta_n^* \le x)$, the conditional distribution function of $\Delta_n^* = m_n(\zeta_n^* - \hat{\zeta}_n)$ and use this to build a CI for ζ_0, where ζ_n^* is the least squares estimator of ζ_0 obtained from the bootstrap sample. We will now introduce four bootstrap schemes that arise naturally in this problem and investigate their consistency properties in Sections 5.4 and 5.5.

Scheme 1 (ECDF bootstrap): Draw a bootstrap sample $(Y_{n,1}^*, Z_{n,1}^*), \ldots, (Y_{n,n}^*, Z_{n,n}^*)$ from the ECDF of $(Y_1, Z_1), \ldots, (Y_n, Z_n)$; probably the most widely used bootstrap scheme.

Scheme 2 (Bootstrapping residuals): This is another widely used bootstrap procedure in regression models. We first obtain the residuals

$$\hat{\epsilon}_{n,j} := Y_j - \hat{\alpha}_n \mathbf{1}_{Z_j \le \hat{\zeta}_n} - \hat{\beta}_n \mathbf{1}_{Z_j > \hat{\zeta}_n} \quad \text{for } j = 1, \ldots, n,$$

from the fitted model. Note that these residuals are not guaranteed to have mean 0, so we work with the centered residuals, $\hat{\epsilon}_{n,1} - \bar{\epsilon}_n, \ldots, \hat{\epsilon}_{n,n} - \bar{\epsilon}_n$, where $\bar{\epsilon}_n = \sum_{j=1}^n \hat{\epsilon}_{n,j}/n$. Letting $\mathbb{P}_n^{\epsilon}$ denote the empirical measure of the centered residuals, we obtain the bootstrap sample $(Y_{n,1}^*, Z_1), \ldots, (Y_{n,n}^*, Z_n)$ as follows:

1. Sample $\epsilon_{n,1}^*, \ldots, \epsilon_{n,n}^*$ independently from $\mathbb{P}_n^{\epsilon}$.
2. Fix the predictors Z_j, $j = 1, \ldots, n$, and define the bootstrapped responses at Z_j as $Y_{n,j}^* = \hat{\alpha}_n \mathbf{1}_{Z_j \le \hat{\zeta}_n} + \hat{\beta}_n \mathbf{1}_{Z_j > \hat{\zeta}_n} + \epsilon_{n,j}^*$.

Scheme 3 (Smoothed bootstrap): Notice that in (5.1) Z is assumed to have a density which arises in the limiting distribution of Δ_n. A successful bootstrap scheme must mimic this underlying assumption, and we accomplish this in the following:

1. Choose an appropriate nonparametric smoothing procedure (e.g., kernel density estimation) to build a distribution $\hat{F}_n$ with a density $\hat{f}_n$ such that $\|\hat{F}_n - F\|_\infty \stackrel{a.s.}{\to} 0$ and $\hat{f}_n \to f$ uniformly on some open interval around ζ_0 w.p. 1, where f is the density of Z.

2. Get i.i.d. replicates $Z^*_{n,1},\ldots,Z^*_{n,n}$ from $\hat{F}_n$ and sample, independently, $\epsilon^*_{n,1},\ldots,\epsilon^*_{n,n}$ from $\mathbb{P}^\epsilon_n$.

3. Define $Y^*_{n,j} = \hat{\alpha}_n \mathbf{1}_{Z^*_{n,j}\le\hat{\zeta}_n} + \hat{\beta}_n \mathbf{1}_{Z^*_{n,j}>\hat{\zeta}_n} + \epsilon^*_{n,j}$ for all $j = 1,\ldots,n$.

Scheme 4 (m out of n bootstrap): A natural alternative to the usual nonparametric bootstrap (i.e., generating bootstrap samples from the ECDF) considered widely in non-regular problems is to use the m out of n bootstrap. We choose a nondecreasing sequence of natural numbers $\{m_n\}_{n=1}^\infty$ such that $m_n = o(n)$ and $m_n \to \infty$ and generate the bootstrap sample $(Y^*_{n,1}, Z^*_{n,1}),\ldots,(Y^*_{n,m_n}, Z^*_{n,m_n})$ from the ECDF of $(Y_1, Z_1),\ldots,(Y_n, Z_n)$. Although there are a number of methods available for choosing the m_n in applications, there is no satisfactory solution to this problem and the obtained CIs usually vary with changing m_n.

We will use the framework established by our convergence theorems in Section 5.3 to prove that schemes 3 and 4 above yield *consistent* bootstrap procedures for building CIs for ζ_0. We will also give strong empirical and theoretical evidence for the *inconsistency* of schemes 1 and 2. Note that schemes 1 and 2 are the two most widely used resampling techniques in regression models (see pages 35-36 of Efron (1982); also see Freedman (1981) and Wu (1986)). Thus in this change–point scenario, a typical nonstandard problem, we see that the two standard bootstrap approaches fail. The failure of the usual bootstrap methods in nonstandard situations is not new and has been investigated in the context of M-estimation problems by Bose and Chatterjee (2001) and in situations giving rise to $n^{1/3}$ asymptotics by Abrevaya and Huang (2005) and Sen et al. (2010). But the change-point problem considered in this paper is indeed quite different from the nonstandard problems considered by the above authors – one key distinction being that compound

Poisson processes, as opposed to Gaussian processes, form the backbone of the asymptotic distributions of the estimators – and thus demands an independent investigation. We will also see later that the performance of scheme 3 clearly dominates that of the m out of n bootstrap procedure (scheme 4), the general recipe proposed in situations where the usual bootstrap does not work (see Lee and Pun (2006) for applications of the m out of n bootstrap procedure in some nonstandard problems). Also note that the performance of the m out of n bootstrap scheme crucially depends on m (see e.g., Bickel et al. (1997)) and the choice of this tuning parameter is tricky in applications.

5.3 A uniform convergence result

In this section we generalize the results obtained in Kosorok (2008b), pages 271–277, to a triangular array of random variables. Consider the triangular array
$\{X_{n,k} = (Y_{n,k}, Z_{n,k})\}_{1\leq k\leq m_n}^{n\in\mathbb{N}}$ defined on a probability space $(\Omega, \mathcal{A}, \mathbf{P})$, where $(m_n)_{n=1}^{\infty}$ is a nondecreasing sequence of natural numbers such that $m_n \to \infty$. Throughout the entire paper we will always denote by $\mathbf{E}$ the expectation operator with respect to $\mathbf{P}$. Furthermore, assume that for each $n \in \mathbb{N}$, $(X_{n,1}, \ldots, X_{n,m_n})$ constitutes a random sample from an arbitrary bivariate distribution $\mathbb{Q}_n$ with $\mathbb{Q}_n(Y_{n,1}^2) < \infty$ and let $M_n(\theta) := \mathbb{Q}_n(m_\theta)$ for all $\theta \in \Theta$, where m_θ is defined in (5.2). Let $\mathbb{P}$ be a bivariate distribution satisfying (5.1). Recall that $M(\theta) := \mathbb{P}(m_\theta)$ and $\theta_0 := \operatorname{sargmax} M(\theta)$.

Let $\theta_n := (\zeta_n, \alpha_n, \beta_n)$ be given by

$$\theta_n = \operatorname*{sargmax}_{\theta\in\Theta}\{\mathbb{Q}_n(m_\theta)\}.$$

Note that $\mathbb{Q}_n$ need not satisfy model (5.1) with $(\zeta_n, \alpha_n, \beta_n)$. The existence of θ_n is guaranteed as $\mathbb{Q}_n(m_\theta)$ is a quadratic function in α and β (for a fixed

ζ) and bounded and cádlág as a function in ζ. For each n, let $\mathbb{P}_n^*$ be the empirical measure produced by the random sample $(X_{n,1},\ldots,X_{n,m_n})$, and define the least squares estimator $\theta_n^* = (\zeta_n^*, \alpha_n^*, \beta_n^*) \in \Theta$ to be the smallest argmax of $M_n^*(\theta) := \mathbb{P}_n^*(m_\theta)$. If Q is a signed Borel measure on $\mathbb{R}^2$ and $\mathscr{F}$ is a class of (possibly) complex-valued functions defined on $\mathbb{R}^2$, write $\|Q\|_{\mathscr{F}} := \sup\{|Q(f)| : f \in \mathscr{F}\}$. If $g : K \subset \mathbb{R}^3 \to \mathbb{R}$ is a bounded function, write $\|g\|_K := \sup_{x\in K}|g(x)|$. Also, for $(z,y) \in \mathbb{R}^2$ and $n \in \mathbb{N}$ we write

$$\tilde{\epsilon}_n := \tilde{\epsilon}_n\,(z,y) = y - \alpha_n \mathbf{1}_{z\le\zeta_n} - \beta_n \mathbf{1}_{z>\zeta_n}. \tag{5.4}$$

Let $M > 0$ be such that $|\alpha_n| \le M$ for all n. We define the following three classes of functions from $\mathbb{R}^2$ into $\mathbb{R}$:

$$\begin{aligned}
\mathcal{F} &:= \{\mathbf{1}_I\,(z) : I \subset \mathbb{R} \text{ is an interval}\}\,,\\
\mathcal{G} &:= \{yf(z) : f \in \mathcal{F}\} \cup \{|y+\alpha| f(z) : f \in \mathcal{F}, |\alpha| \le M\}\,,\\
\mathcal{H} &:= \{y^2 f(z) : f \in \mathcal{F}\}.
\end{aligned}$$

In what follows, we will derive conditions on the distributions $\mathbb{Q}_n$ that will guarantee consistency and weak convergence of θ_n^*.

5.3.1 Consistency and the rate of convergence

We provide first a consistency result for the least squares estimator, whose proof we include in the Appendix (see Section 5.9.1). To this end, we consider the following set of assumptions:

(I) $\|\mathbb{Q}_n - \mathbb{P}\|_{\mathcal{F}} \to 0$,

(II) $\|\mathbb{Q}_n - \mathbb{P}\|_{\mathcal{G}} \to 0$,

(III) $\|\mathbb{Q}_n - \mathbb{P}\|_{\mathcal{H}} \to 0$,

(IV) $\theta_n \to \theta_0$.

Proposition 5.3.1 *Assume that (I)-(IV) hold. Then,* $\theta_n^* \xrightarrow{\mathbf{P}} \theta_0$.

To guarantee the right rate of convergence, we need to assume stronger regularity conditions. In addition to those of Proposition 5.3.1, we require the following:

(V) There are $\eta, \rho, L > 0$ with the property that for any $\delta \in (0, \eta)$, there is $N > 0$ such that the following inequalities hold for any $n \geq N$:

$$\inf_{\frac{1}{\sqrt{m_n}} \leq |\zeta - \zeta_n| < \delta^2} \left\{ \frac{1}{|\zeta - \zeta_n|} \mathbb{Q}_n(\mathbf{1}_{\zeta \wedge \zeta_n < Z \leq \zeta \vee \zeta_n}) \right\} > \rho, \tag{5.5}$$

$$\sup_{|\zeta - \zeta_n| < \delta^2} \{ |\mathbb{Q}_n(\tilde{\epsilon}_n \mathbf{1}_{\zeta \wedge \zeta_n < Z \leq \zeta \vee \zeta_n})| \} \leq \frac{L\delta}{\sqrt{m_n}}, \tag{5.6}$$

$$\sup_{|\zeta - \zeta_n| < \delta^2} \{ |\mathbb{Q}_n(\tilde{\epsilon}_n \mathbf{1}_{Z \leq \zeta \wedge \zeta_n})| + |\mathbb{Q}_n(\tilde{\epsilon}_n \mathbf{1}_{Z > \zeta \vee \zeta_n})| \} \leq \frac{L}{\sqrt{m_n}}. \tag{5.7}$$

We would like to point out some facts about (V). It must be noted that (5.6) and (5.7) automatically hold in the case where Z and $\tilde{\epsilon}_n$ are independent under $\mathbb{Q}_n$ with $\mathbb{Q}_n(\tilde{\epsilon}_n) = 0$. Also, (5.5) is easily seen to hold when the Z's, under $\mathbb{Q}_n$, have densities f_n converging uniformly to f in some neighborhood of ζ_0, where f is the density of Z under $\mathbb{P}$; by a consequence of the classical mean value theorem of calculus.

With the aid of these conditions, Proposition 5.3.1 and Theorem 3.4.1, page 322, of van der Vaart and Wellner (1996) we can now state and prove (see Section 5.9.2) the rate of convergence result.

Proposition 5.3.2 *Assume that (I)-(V) hold. Then* $\sqrt{m_n}(\alpha_n^* - \alpha_n) = O_{\mathbf{P}}(1)$, $\sqrt{m_n}(\beta_n^* - \beta_n) = O_{\mathbf{P}}(1)$ *and* $m_n(\zeta_n^* - \zeta_n) = O_{\mathbf{P}}(1)$.

Propositions 5.3.1 and 5.3.2 provide sufficient conditions on the measures $\mathbb{Q}_n$, the distribution of each element in the nth row of the triangular array, to achieve the same rate of convergence as the original least squares estimators. We would like to highlight that we are not assuming that each $\mathbb{Q}_n$ satisfy the model (5.1) with $(\alpha_n, \beta_n, \zeta_n)$; all we need is that $\mathbb{Q}_n$ and θ_n approach $\mathbb{P}$ and θ_0 respectively, in a suitable manner.

5.3.2 Weak Convergence and asymptotic distribution

We start with some additional set of assumptions:

(VI) For any function $\psi : \mathbb{R} \to \mathbb{C}$ which is either of the form $\psi(x) = e^{i\xi x}$ for some $\xi \in \mathbb{R}$ or defined by $\psi(x) = |x|^p$ for $p = 1, 2$, we have:

$$m_n \mathbb{Q}_n \left(\psi(\tilde{\epsilon}_n) \mathbf{1}_{\zeta_n - \frac{\delta}{m_n} < Z \le \zeta_n + \frac{\eta}{m_n}} \right) \to f(\zeta_0)(\delta + \eta) \mathbb{P}\left(\psi(\epsilon)\right) \quad \forall\, \eta, \delta > 0.$$

(VII) $\sqrt{m_n}\mathbb{Q}_n(\tilde{\epsilon}_n \mathbf{1}_{Z \le \zeta_n}) \to 0$ and $\sqrt{m_n}\mathbb{Q}_n(\tilde{\epsilon}_n \mathbf{1}_{Z > \zeta_n}) \to 0$.

(VIII) $\overline{\lim}_{n \to \infty} \mathbb{Q}_n(|\tilde{\epsilon}_n|^3) < \infty$.

Observe that condition (VI) implies, for all $\eta, \delta > 0$, and $p = 1, 2$,

$$\sqrt{m_n}\mathbb{Q}_n \left(|\tilde{\epsilon}_n|^p \mathbf{1}_{\zeta_n - \frac{\delta}{m_n} < Z \le \zeta_n + \frac{\eta}{m_n}} \right) \to 0, \tag{5.8}$$

$$\sqrt{m_n}\mathbb{Q}_n \left(\mathbf{1}_{\zeta_n - \frac{\delta}{m_n} < Z \le \zeta_n + \frac{\eta}{m_n}} \right) \to 0. \tag{5.9}$$

For $h = (h_1, h_2, h_3) \in \mathbb{R}^3$, let $\vartheta_{n,h} := \theta_n + \left(\frac{h_1}{m_n}, \frac{h_2}{\sqrt{m_n}}, \frac{h_3}{\sqrt{m_n}} \right)$ and

$$\hat{E}_n(h) := m_n \mathbb{P}_n^* \left[m_{\vartheta_{n,h}} - m_{\theta_n} \right].$$

We will argue that

$$h_n^* := \underset{h \in \mathbb{R}^3}{\mathrm{sargmin}}\, \hat{E}_n(h) = (m_n(\zeta_n^* - \zeta_n), \sqrt{m_n}(\alpha_n^* - \alpha_n), \sqrt{m_n}(\beta_n^* - \beta_n))$$

converges in distribution to the smallest argmax of some process involving two independent normal random variables and a two-sided, compound Poisson process (independent of the normal variables).

We will derive the asymptotic distribution of the process $\hat{E}_n$ and then apply Theorem 4.3.2 to obtain the limiting distribution of h_n^*. We will consider these stochastic processes as random elements in the Skorohod spaces $\mathcal{D}_K$ as given Definition 4.2.2. As a first step in this direction, we express the process $\hat{E}_n$ as the sum of the four terms $\hat{A}_n$, $\hat{B}_n$, $\hat{C}_n$ and $\hat{D}_n$ where

$$\begin{aligned}
\hat{A}_n(h_1,h_2) &:= 2h_2\sqrt{m_n}\mathbb{P}_n^*\left(\tilde{\epsilon}_n\mathbf{1}_{Z\le\zeta_n\wedge\left(\zeta_n+\frac{h_1}{m_n}\right)}\right) - h_2^2\mathbb{P}_n^*\left(\mathbf{1}_{Z\le\zeta_n\wedge\left(\zeta_n+\frac{h_1}{m_n}\right)}\right),\\
\hat{B}_n(h_1,h_3) &:= 2h_3\sqrt{m_n}\mathbb{P}_n^*\left(\tilde{\epsilon}_n\mathbf{1}_{Z>\zeta_n\vee\left(\zeta_n+\frac{h_1}{m_n}\right)}\right) - h_3^2\mathbb{P}_n^*\left(\mathbf{1}_{Z>\zeta_n\vee\left(\zeta_n+\frac{h_1}{m_n}\right)}\right),\\
\hat{C}_n(h_1,h_3) &:= -2m_n\left(\alpha_n-\beta_n-\frac{h_3}{\sqrt{m_n}}\right)\mathbb{P}_n^*\left(\tilde{\epsilon}_n\mathbf{1}_{\zeta_n+\frac{h_1}{m_n}<Z\le\zeta_n}\right)\\
&\qquad - m_n\left(\alpha_n-\beta_n-\frac{h_3}{\sqrt{m_n}}\right)^2\mathbb{P}_n^*\left(\mathbf{1}_{\zeta_n+\frac{h_1}{m_n}<Z\le\zeta_n}\right),\\
\hat{D}_n(h_1,h_2) &:= -2m_n\left(\beta_n-\alpha_n-\frac{h_2}{\sqrt{m_n}}\right)\mathbb{P}_n^*\left(\tilde{\epsilon}_n\mathbf{1}_{\zeta_n<Z\le\zeta_n+\frac{h_1}{m_n}}\right)\\
&\qquad - m_n\left(\beta_n-\alpha_n-\frac{h_2}{\sqrt{m_n}}\right)^2\mathbb{P}_n^*\left(\mathbf{1}_{\zeta_n<Z\le\zeta_n+\frac{h_1}{m_n}}\right).
\end{aligned}$$

We define another process $E_n^* := A_n^* + B_n^* + C_n^* + D_n^*$ where

$$\begin{aligned}
A_n^*(h_2) &:= 2h_2\sqrt{m_n}\mathbb{P}_n^*\left(\tilde{\epsilon}_n\mathbf{1}_{Z\le\zeta_n}\right) - h_2^2\mathbb{P}_n^*\left(\mathbf{1}_{Z\le\zeta_n}\right),\\
B_n^*(h_3) &:= 2h_3\sqrt{m_n}\mathbb{P}_n^*\left(\tilde{\epsilon}_n\mathbf{1}_{Z>\zeta_n}\right) - h_3^2\mathbb{P}_n^*\left(\mathbf{1}_{Z>\zeta_n}\right),\\
C_n^*(h_1) &:= -2m_n(\alpha_n-\beta_n)\mathbb{P}_n^*\left(\tilde{\epsilon}_n\mathbf{1}_{\zeta_n+\frac{h_1}{m_n}<Z\le\zeta_n}\right)\\
&\qquad - m_n(\alpha_n-\beta_n)^2\mathbb{P}_n^*\left(\mathbf{1}_{\zeta_n+\frac{h_1}{m_n}<Z\le\zeta_n}\right),\\
D_n^*(h_1) &:= -2m_n(\beta_n-\alpha_n)\mathbb{P}_n^*\left(\tilde{\epsilon}_n\mathbf{1}_{\zeta_n<Z\le\zeta_n+\frac{h_1}{m_n}}\right)\\
&\qquad - m_n(\beta_n-\alpha_n)^2\mathbb{P}_n^*\left(\mathbf{1}_{\zeta_n<Z\le\zeta_n+\frac{h_1}{m_n}}\right).
\end{aligned}$$

We work with E_n^* instead of $\hat{E}_n$ as their difference approaches uniformly to 0 in probability, as shown in the next lemma (proved in Section 5.9.3), and the asymptotic distribution of E_n^* is easier to derive.

Lemma 5.3.1 *Let $K \subset \mathbb{R}^3$ be a compact rectangle. If conditions (I)-(IV) and (5.8) and (5.9) hold, then*

$$\left\| E_n^* - \hat{E}_n \right\|_K \xrightarrow{\mathbf{P}} 0.$$

Therefore, $E_n^* - \hat{E}_n \xrightarrow{\mathbf{P}} 0$ as random elements of $\mathcal{D}_K$. In particular, this result is true under conditions (I)-(IV) and (VI).

As a first step to finding the asymptotic distribution of $(E_n^*)_{n=1}^{\infty}$, we show that the random sequence is tight in the Skorohod space $\mathcal{D}_K$ for any compact rectangle $K \subset \mathbb{R}^3$. The proof of the next result is given in Section 5.9.4.

Lemma 5.3.2 *Let $I \subset \mathbb{R}$ be a compact interval and assume that conditions (I)-(VIII) hold. Then, the sequence of $\mathbb{R}^6$-valued processes*

$$\Xi_n(t) := \begin{pmatrix} \sqrt{m_n}\mathbb{P}_n^*(\tilde{\epsilon}_n \mathbf{1}_{Z \leq \zeta_n}) \\ \sqrt{m_n}\mathbb{P}_n^*(\tilde{\epsilon}_n \mathbf{1}_{Z > \zeta_n}) \\ m_n\mathbb{P}_n^*(\mathbf{1}_{\zeta_n + \frac{t}{m_n} < Z \leq \zeta_n}) \\ m_n\mathbb{P}_n^*(\tilde{\epsilon}_n \mathbf{1}_{\zeta_n + \frac{t}{m_n} < Z \leq \zeta_n}) \\ m_n\mathbb{P}_n^*(\mathbf{1}_{\zeta_n < Z \leq \zeta_n + \frac{t}{m_n}}) \\ m_n\mathbb{P}_n^*(\tilde{\epsilon}_n \mathbf{1}_{\zeta_n < Z \leq \zeta_n + \frac{t}{m_n}}) \end{pmatrix} \tag{5.10}$$

is uniformly tight in $\mathbb{R}^2 \times \tilde{\mathcal{D}_I^4}$. Also, if $K \subset \mathbb{R}^3$ is a compact rectangle, the sequence $(E_n^*)_{n=1}^{\infty}$ is uniformly tight in $\mathcal{D}_K$.

It now suffices to show convergence of the finite-dimensional distributions of the processes E_n^* to the finite dimensional distributions of some process $E^* \in \mathcal{D}_K$ to conclude that E_n^* converges weakly to E^* (and thus $\hat{E}_n$ too). With this objective in mind, we make the following definitions: Let $\mathbf{Z}_1 \sim \mathbf{N}\left(0, \sigma^2\mathbb{P}(Z \leq \zeta_0)\right)$ and $\mathbf{Z}_2 \sim \mathbf{N}\left(0, \sigma^2\mathbb{P}(Z > \zeta_0)\right)$ be two independent normal random variables; ν_1 and ν_2 be, respectively, left-continuous and right-continuous, homogeneous Poisson processes with rate $f(\zeta_0) > 0$; $\mathbf{u} = (u_n)_{n=1}^{\infty}$ and $\mathbf{v} = (v_n)_{n=1}^{\infty}$ two sequences of i.i.d. random variables having the same distribution as ϵ under $\mathbb{P}$. Assume, in addition, that $\mathbf{Z}_1$, $\mathbf{Z}_2$, ν_1, ν_2, $\mathbf{v}$ and $\mathbf{u}$ are all mutually independent. Then, define the process $\Xi = (\Xi^{(1)}, \dots, \Xi^{(6)})'$ as

$$\Xi(t) := \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \\ \nu_1(-t)\mathbf{1}_{t<0} \\ \sum_{0<j\leq \nu_1(-t)} v_j \mathbf{1}_{t<0} \\ \nu_2(t)\mathbf{1}_{t\geq 0} \\ \sum_{0<j\leq \nu_2(t)} u_j \mathbf{1}_{t\geq 0} \end{pmatrix} \tag{5.11}$$

and let E^* be given by

$$\begin{aligned} E^*(h) \;:=\; & 2h_2\Xi^{(1)}(h_1) - h_2^2\mathbb{P}(Z \leq \zeta_0) + 2h_3\Xi^{(2)}(h_1) - h_3^2\mathbb{P}(Z > \zeta_0) \\ & + 2(\beta_0 - \alpha_0)\Xi^{(4)}(h_1) - (\alpha_0 - \beta_0)^2\Xi^{(3)}(h_1) \\ & + 2(\alpha_0 - \beta_0)\Xi^{(6)}(h_1) - (\alpha_0 - \beta_0)^2\Xi^{(5)}(h_1) \end{aligned} \tag{5.12}$$

for $h = (h_1, h_2, h_3) \in \mathbb{R}^3$.

We will now prove weak convergence of the sequence of processes $(\hat{E}_n)_{n=1}^{\infty}$ to E^* and apply Theorem 4.3.2. Recall the notation of Section 4.3 and define the $\mathcal{S}$–valued (pure jump) processes $\hat{J}_n$, J_n^* and J^* as

$$\begin{aligned} J_n^*(t) = \hat{J}_n(t) \;&:=\; m_n\mathbb{P}_n^*(\mathbf{1}_{\zeta_n + \frac{t}{m_n} < Z \leq \zeta_n}) + m_n\mathbb{P}_n^*(\mathbf{1}_{\zeta_n < Z \leq \zeta_n + \frac{t}{m_n}}), \\ J^*(t) \;&:=\; \nu_1(-t)\mathbf{1}_{t<0} + \nu_2(t)\mathbf{1}_{t\geq 0}. \end{aligned}$$

Note that $\hat{J}_n$, J_n^* and J^* are the jump processes associated with $\hat{E}_n$, E_n^* and E^*, respectively.

Lemma 5.3.3 *Let $I \subset \mathbb{R}$ be a compact interval and $K = I \times A \times B \subset \mathbb{R}^3$ a compact rectangle. If (I)-(VIII) hold, we have*

(i) $\Xi_n \rightsquigarrow \Xi$ *in* $\mathbb{R}^2 \times \tilde{\mathcal{D}}_I^4$,

(ii) $(E_n^*, J_n^*) \rightsquigarrow (E^*, J^*)$ *in* $\mathcal{D}_K \times \mathcal{S}_I$,

(iii) $(\hat{E}_n, \hat{J}_n) \rightsquigarrow (E^*, J^*)$ *in* $\mathcal{D}_K \times \mathcal{S}_I$,

where $\rightsquigarrow$ denotes weak convergence.

For a proof of the convergence result, see Section 5.9.5.

To apply Theorem 4.3.2 we first show that the the smallest argmax of E^* is well defined. The proof of the next lemma is provided in Section 5.9.6.

Lemma 5.3.4 *Consider the process E^* defined in (5.12). Then, for almost every sample path of E^*, $\phi^* = (\phi_1^*, \phi_2^*, \phi_3^*) := \underset{h \in \mathbb{R}^3}{sargmax}\{E^*(h)\}$ is well-defined. Moreover, ϕ_1^*, ϕ_2^* and ϕ_3^* are independent; and ϕ_2^* and ϕ_3^* are distributed as normal random variables with mean 0 and variances $\sigma^2/\mathbb{P}(Z \leq \zeta_0)$ and $\sigma^2/\mathbb{P}(Z > \zeta_0)$, respectively.*

We now state the distributional convergence result for the sequence of least squares estimator θ_n^*. For a proof, we refer the reader to Section 5.9.7.

Proposition 5.3.3 *With the notation of Lemma 5.3.4, if conditions (I)-(VIII) hold, then*

$$h_n^* = \begin{pmatrix} m_n(\zeta_n^* - \zeta_n) \\ \sqrt{m_n}(\alpha_n^* - \alpha_n) \\ \sqrt{m_n}(\beta_n^* - \beta_n) \end{pmatrix} \rightsquigarrow \underset{h \in \mathbb{R}^3}{sargmax}\{E^*(h)\}.$$

If we take $\mathbb{Q}_n = \mathbb{P}$ and $m_n = n \; \forall n \in \mathbb{N}$, it is easily seen that $\theta_n = \theta_0$ and conditions (I)-(VIII) hold. Hence, we immediately get the following corollary.

Corollary 5.3.1 (Asymptotic distribution of the least squares estimators) *For the least squares estimators* $(\hat{\zeta}_n, \hat{\alpha}_n, \hat{\beta}_n)$ *based on an i.i.d. sequence* $(X_n)_{n=1}^{\infty}$ *satisfying (5.1), we have*

$$(n(\hat{\zeta}_n - \zeta_0), \sqrt{n}(\hat{\alpha}_n - \alpha_0), \sqrt{n}(\hat{\beta}_n - \beta_0))' \rightsquigarrow \underset{h \in \mathbb{R}^3}{\operatorname{sargmax}}\{E^*(h)\}.$$

5.4 Inconsistency of the bootstrap

In this section we argue the inconsistency of the two most common bootstrap procedures in regression: the ECDF bootstrap (scheme 1) and the residual bootstrap (scheme 2). Recall the notation and definitions in the beginning of Section 5.2. In particular, note that we have i.i.d. random vectors $\{X_n = (Y_n, Z_n)\}_{n=1}^{\infty}$ from (5.1) with parameter θ_0 defined on a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ and let $\mathbb{P}_n$ be the empirical distribution of the first n data points. We start by stating two results that will be used in the sequel. We first show that the least squares estimator $\hat{\theta}_n$ of θ_0 is strongly consistent. This is an improvement of the result obtained in Kosorok (2008b) and we refer the reader to Section 5.9.8 for a complete proof. The proof of the second lemma can be found in Section 5.9.9.

Lemma 5.4.1 *Let* $K \subset \Theta$ *be any compact rectangle. Then,*

(i) $\|M_n - M\|_K \xrightarrow{a.s.} 0$,

(ii) $M_n \xrightarrow{a.s.} M$ *in* $\mathcal{D}_K$,

(iii) $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$.

Lemma 5.4.2 *Let* $K \subset \mathbb{R}$ *be a compact interval and* $(m_n)_{n=1}^{\infty}$ *be an increasing sequence of natural numbers such that* $m_n \to \infty$ *and* $m_n = O(n)$. *Then,*

(i) $m_n^{\gamma} \left\| \mathbb{P}_n(\hat{\zeta}_n + \frac{(\cdot)}{m_n} < Z \leq \hat{\zeta}_n) \right\|_K \xrightarrow{\mathbf{P}} 0$ *for any* $\gamma < 1$, *and*

(ii) $m_n^\gamma \left\| \mathbb{P}_n \left(|\tilde{\epsilon}_n|^p \mathbf{1}_{\hat{\zeta}_n + \frac{(\cdot)}{m_n} < Z \leq \hat{\zeta}_n} \right) \right\|_K \xrightarrow{\mathbf{P}} 0$ *for any* $\gamma < 1$, *and* $p = 1{,}2$.

These statements are still true if $\mathbf{1}_{\hat{\zeta}_n + \frac{(\cdot)}{m_n} < Z \leq \hat{\zeta}_n}$ *is replaced by* $\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{(\cdot)}{m_n}}$.

We introduce some notation. Let $(\mathtt{X}, d)$ be a metric space and consider the $\mathtt{X}$-valued random elements V and $(V_n)_{n=1}^\infty$ defined on $(\Omega, \mathcal{A}, \mathbf{P})$. We say that V_n converges conditionally in probability to V, almost surely, and write $V_n \overset{\mathbf{P}_{\mathfrak{X}}}{\underset{a.s.}{\to}} V$, if

$$\mathbf{P}_{\mathfrak{X}}(d(V_n, V) > \epsilon) \xrightarrow{a.s.} 0 \quad \forall\, \epsilon > 0. \tag{5.13}$$

Similarly, we write $V_n \overset{\mathbf{P}_{\mathfrak{X}}}{\underset{\mathbf{P}}{\to}} V$ and say that V_n converges conditionally in probability to V, in probability, if the left–hand side of (5.13) converges in probability to 0.

5.4.1 Scheme 1 (Bootstrapping from ECDF)

Consider the notation and definitions of Section 5.2.1. To translate this scheme into the framework of Propositions 5.3.1, 5.3.2 and 5.3.3, we set $m_n = n$, $\mathbb{Q}_n = \mathbb{P}_n$ and consider the triangular array $\left\{X_{n,k}^* = (Y_{n,k}^*, Z_{n,k}^*)\right\}_{1 \leq k \leq n}^{n \in \mathbb{N}}$. Moreover, from Lemma 5.4.1 we know that $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$, so we can also take $\theta_n = \hat{\theta}_n$. We first prove that the bootstrapped estimators converge conditionally in probability to the true value of the parameters, almost surely.

Proposition 5.4.1 *For the ECDF bootstrap, we have* $\theta_n^* \overset{\mathbf{P}_{\mathfrak{X}}}{\underset{a.s.}{\to}} \theta_0$.

Proof: Since Y has a second moment under $\mathbb{P}$, it is straightforward to see that $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{H}$ are VC-subgraph classes with integrable envelopes 1, $|Y| + M$ and Y^2, respectively. It follows that all these classes are Glivenko–Cantelli and therefore conditions (I)-(III) hold w.p. 1. Also, note that, from Lemma 5.4.1 *(iii)* condition (IV) holds a.s. The result then follows from

Proposition 5.3.1. □

Let $\mathbb{P}_n^*$ be the ECDF of $X_{n,1}^*, \ldots, X_{n,n}^*$ and recall the definition of the processes $\hat{A}_n$, $\hat{B}_n$, $\hat{C}_n$, $\hat{D}_n$, $\hat{E}_n$, A_n^*, B_n^*, C_n^*, D_n^* and E_n^*. We then have the following result.

Lemma 5.4.3 *Let $K \subset \mathbb{R}^3$ be any compact rectangle. Then*

$$\hat{E}_n - E_n^* \xrightarrow[\mathbf{P}]{\mathbf{P}_{\mathfrak{X}}} 0 \text{ in } \mathcal{D}_K.$$

Proof: We already know that conditions (I)-(IV) hold w.p. 1 under this bootstrap scheme. But Lemma 5.4.2 implies that (5.8) and (5.9) hold in probability. Hence, this result follows by arguing through subsequences and applying Lemma 5.3.1. □

It is evident that condition (VI) doesn't hold in this situation as we know that

$$n\mathbb{P}_n\left(\zeta_0 - \frac{\eta}{n} < Z \leq \zeta_0 + \frac{\delta}{n}\right) \rightsquigarrow \text{Poisson}\big(f(\zeta_0)(\delta + \eta)\big). \tag{5.14}$$

Hence, we cannot use Proposition 5.3.3 to derive the limit behavior of h_n^*.

We will now argue that E_n^*, and therefore $\hat{E}_n$, *does not have any weak limit* in probability. This statement should be thought in terms of the Prokhorov metric (or any other metric metrizing weak convergence on $\mathcal{D}_K$). If we denote by ρ_K the Prokhorov metric on the space of probability measures on $\mathcal{D}_K$ and by μ_n the conditional distribution of E_n^* given $\mathfrak{X}$, to say that $(E_n^*)_{n=1}^{\infty}$ has no weak limit in probability means that there is no probability measure μ defined on $\mathcal{D}_K$ such that $\rho_K(\mu_n, \mu) \xrightarrow{\mathbf{P}} 0$.

The following lemma (proved in Section 5.9.10) will help us show that the (conditional) characteristic functions corresponding to the finite dimensional distributions of E_n^* fail to have a limit in probability, which would, in particular, imply that E_n^* does not have a weak limit in probability.

Lemma 5.4.4 *The following statements hold:*

(i) For any two real numbers $s < t$, $\left\{n\mathbb{P}_n(\zeta_0 + \frac{s}{n} < Z \leq \zeta_0 + \frac{t}{n})\right\}_{n=1}^{\infty}$ does not converge in probability.

(ii) There is $h_* > 0$ such that for any $h \geq h_*$, the sequences $\left\{n\mathbb{P}_n(\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{h}{n})\right\}_{n=1}^{\infty}$ and $\left\{n\mathbb{P}_n(\hat{\zeta}_n - \frac{h}{n} < Z \leq \hat{\zeta}_n)\right\}_{n=1}^{\infty}$ do not converge in probability.

(iii) For any two real numbers $s < t$ and any measurable function $\phi : \mathbb{R} \to \mathbb{R}$, $\left\{n\mathbb{P}_n(\phi(Y)\mathbf{1}_{\zeta_0+\frac{s}{n}<Z\leq\zeta_0+\frac{t}{n}})\right\}_{n=1}^{\infty}$ does not converge in probability.

(iv) Let ϕ be a measurable function which is either nonnegative or nonpositive and such that $\phi(\epsilon + \alpha_0)$ and $\phi(\epsilon + \beta_0)$ are nonconstant random variables with finite second moment. Then, there is $h_* > 0$ such that for any $h \geq h_*$ $\left\{n\mathbb{P}_n(\phi(Y)\mathbf{1}_{\hat{\zeta}_n<Z\leq\hat{\zeta}_n+\frac{h}{n}})\right\}_{n=1}^{\infty}$ and $\left\{n\mathbb{P}_n(\phi(Y)\mathbf{1}_{\hat{\zeta}_n-\frac{h}{n}<Z\leq\hat{\zeta}_n})\right\}_{n=1}^{\infty}$ do not converge in probability.

With the aid of Lemma 5.4.4 we are now able to state our main result.

Lemma 5.4.5 *There is a compact rectangle $K \subset \mathbb{R}^3$ such that neither $\hat{E}_n$ nor E_n^* has a weak limit in probability in $\mathcal{D}_K$.*

Proof: Since Lemma 5.4.3 and Slutsky's lemma show that $\hat{E}_n$ has a weak limit in probability if and only if E_n^* has a weak limit in probability, it suffices to argue that the statement is true for E_n^*. To prove this, it is enough to show that there is some h_1 such that $E_n^*(h_1, 0, 0)$ does not converge in distribution. Pick $h_1 > 0$ and observe that

$$E_n^*(h_1, 0, 0) = (\hat{\alpha}_n - \hat{\beta}_n)\left(n\mathbb{P}_n^*\left[(2\tilde{\epsilon}_n - \hat{\alpha}_n + \hat{\beta}_n)\mathbf{1}_{\hat{\zeta}_n<Z\leq\hat{\zeta}_n+\frac{h_1}{n}}\right]\right).$$

Since $\hat{\alpha}_n - \hat{\beta}_n \xrightarrow{a.s.} \alpha_0 - \beta_0 \neq 0$ we see that $E_n^*(h_1, 0, 0)$ will converge weakly in probability if and only if $\Lambda_n := n\mathbb{P}_n^* \left[(2\tilde{\epsilon}_n - \hat{\alpha}_n + \hat{\beta}_n)\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{h_1}{n}} \right]$ converges weakly in probability.

The conditional characteristic function of Λ_n is given by

$$\mathbf{E}_{\mathfrak{X}}\left(e^{i\xi\Lambda_n}\right) = \left(1 + \frac{1}{n} n\mathbb{P}_n \left((\mathrm{e}^{i\xi(2\tilde{\epsilon}_n + \hat{\beta}_n - \hat{\alpha}_n)} - 1)\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{h_1}{n}} \right)\right)^n, \quad (5.15)$$

which converges in probability if and only if so does

$$n\mathbb{P}_n \left((\mathrm{e}^{i\xi(2\tilde{\epsilon}_n + \hat{\beta}_n - \hat{\alpha}_n)} - 1)\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{h_1}{n}} \right).$$

. But note that

$$n\mathbb{P}_n \left((\mathrm{e}^{i\xi(2\tilde{\epsilon}_n + \hat{\beta}_n - \hat{\alpha}_n)} - 1)\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{h_1}{n}} \right) = n\mathbb{P}_n \left((\mathrm{e}^{i\xi(2Y - \hat{\beta}_n - \hat{\alpha}_n)} - 1)\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{h_1}{n}} \right).$$

It is easily seen that (5.14) and the fact that $n(\hat{\zeta}_n - \zeta_0) = O_{\mathbf{P}}(1)$ imply that

$$n\mathbb{P}_n \left(\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{h_1}{n}} \right) = O_{\mathbf{P}}(1).$$

Hence,

$$\left| n\mathbb{P}_n \left((\mathrm{e}^{i\xi(2Y - \hat{\beta}_n - \hat{\alpha}_n)} - 1)\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{h_1}{n}} \right) - n\mathbb{P}_n \left((\mathrm{e}^{i\xi(2Y - \beta_0 - \alpha_0)} - 1)\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{h_1}{n}} \right) \right|$$
$$\leq n\mathbb{P}_n \left(\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{h_1}{n}} \right) (|\hat{\alpha}_n - \alpha_0| + |\hat{\beta}_n - \beta_0|)|\xi| \xrightarrow{\mathbf{P}} 0.$$

It follows that $\mathbf{E}_{\mathfrak{X}}\left(e^{i\xi\Lambda_n}\right)$ has a limit in probability if and only if

$$n\mathbb{P}_n \left((\mathrm{e}^{i\xi(2Y - \beta_0 - \alpha_0)} - 1)\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{h_1}{n}} \right)$$

has a limit in probability. But a necessary condition for the latter to happen is that its real part,

$$n\mathbb{P}_n \left(\mathrm{Re}(\mathrm{e}^{i\xi(2Y - \beta_0 - \alpha_0)} - 1)\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{h_1}{n}} \right)$$

converges in probability. Since $\mathrm{Re}(\mathrm{e}^{i\xi(2Y - \beta_0 - \alpha_0)} - 1) \leq 0$ we can conclude from (iv) of Lemma 5.4.4 that $n\mathbb{P}_n \left(\mathrm{Re}(\mathrm{e}^{i\xi(2Y - \beta_0 - \alpha_0)} - 1)\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{h_1}{n}} \right)$ does not converge in probability for all $h_1 \geq h_*$ for some $h_* > 0$ large enough. This in turn implies that, for all $h_1 \geq h_*$, the conditional characteristic function in (5.15) does not converge in probability and hence $E_n^*(h_1, 0, 0)$ has no weak limit in probability.

Hence, if K is any compact rectangle containing $(h_*, 0, 0)$ the finite dimensional dimensional distributions of E_n^* on K do not have a weak limit in probability. Therefore, E_n^* does not have a weak limit in probability on $\mathcal{D}_K$. $\square$

Note that

$$\left(n(\zeta_n^* - \hat{\zeta}_n), \sqrt{n}(\alpha_n^* - \hat{\alpha}_n), \sqrt{n}(\beta_n^* - \hat{\beta}_n)\right) = \operatorname*{sargmax}_{h \in \mathbb{R}^3} \left\{\hat{E}_n(h)\right\}.$$

Thus, the fact that the sequence $(\hat{E}_n)_{n=1}^{\infty}$ doesn't have a weak limit in probability makes the existence of a weak limit in probability for $n(\zeta_n^* - \hat{\zeta}_n)$ very unlikely. However, we do not have a rigorous mathematical proof of this statement. The main difficulty in such a proof is that the sargmax functional is non-linear and that $\hat{E}_n$ depends on h_3 through indicator functions that do not converge in the limit.

Remark: It must be noted in this connection that the bootstrap scheme estimates the distribution of $(\sqrt{n}(\alpha_n^* - \hat{\alpha}_n), \sqrt{n}(\beta_n^* - \hat{\beta}_n))$ correctly, and in fact, valid bootstrap based inference can be conducted to obtain CIs for α_0 and β_0. This follows from the fact that, asymptotically, the maximizers of $\hat{E}_n(h_1, \cdot, \cdot)$ do not depend on h_1 (see the expressions for $\hat{A}_n$, $\hat{B}_n$, A_n^*, B_n^*).

We next provide an alternative additional argument that illustrates the inconsistency of the ECDF bootstrap. Our approach is similar to that of Kosorok (2008a) and relies on the asymptotic *unconditional* behavior of

$$\tilde{\Delta}_n^* := (n(\zeta_n^* - \zeta_0), \sqrt{n}(\alpha_n^* - \alpha_0), \sqrt{n}(\beta_n^* - \beta_0)).$$

For $h \in \mathbb{R}^3$, we write $\tilde{\vartheta}_{n,h} := \theta_0 + \left(\frac{h_1}{n}, \frac{h_2}{\sqrt{n}}, \frac{h_3}{\sqrt{n}}\right)$ and

$$\tilde{E}_n(h) := n\mathbb{P}_n^*\left[m_{\tilde{\vartheta}_{n,h}} - m_{\theta_0}\right]. \tag{5.16}$$

This corresponds to centering the objective function around θ_0. As in (5.10), we can define the processes

$$\tilde{\Xi}_n(t) = \begin{pmatrix} \tilde{\Xi}_n^{(1)}(t) \\ \tilde{\Xi}_n^{(2)}(t) \\ \tilde{\Xi}_n^{(3)}(t) \\ \tilde{\Xi}_n^{(4)}(t) \\ \tilde{\Xi}_n^{(5)}(t) \\ \tilde{\Xi}_n^{(6)}(t) \end{pmatrix} := \begin{pmatrix} \sqrt{n}\mathbb{P}_n^*(\epsilon \mathbf{1}_{Z \le \zeta_0}) \\ \sqrt{n}\mathbb{P}_n^*(\epsilon_n \mathbf{1}_{Z > \zeta_0}) \\ n\mathbb{P}_n^*(\mathbf{1}_{\zeta_0 + \frac{t}{n} < Z \le \zeta_0}) \\ n\mathbb{P}_n^*(\epsilon \mathbf{1}_{\zeta_0 + \frac{t}{n} < Z \le \zeta_0}) \\ n\mathbb{P}_n^*(\mathbf{1}_{\zeta_0 < Z \le \zeta_0 + \frac{t}{n}}) \\ n\mathbb{P}_n^*(\epsilon \mathbf{1}_{\zeta_0 < Z \le \zeta_0 + \frac{t}{n}}) \end{pmatrix} \tag{5.17}$$

and just as in that case, we can also define the process $\tilde{E}_n^*$ by

$$\begin{aligned} \tilde{E}_n^*(h) \;\; := \;\; & 2h_2\tilde{\Xi}_n^{(1)}(h_1) - h_2^2\mathbb{P}_n^*(Z \le \zeta_0) + 2h_3\tilde{\Xi}_n^{(2)}(h_1) - h_3^2\mathbb{P}_n^*(Z > \zeta_0) \\ & + 2(\beta_0 - \alpha_0)\tilde{\Xi}_n^{(4)}(h_1) - (\alpha_0 - \beta_0)^2\tilde{\Xi}_n^{(3)}(h_1) \\ & + 2(\alpha_0 - \beta_0)\tilde{\Xi}_n^{(6)}(h_1) - (\alpha_0 - \beta_0)^2\Xi_n^{(5)}(h_1). \end{aligned}$$

Then, it can be shown that $\tilde{E}_n - \tilde{E}_n^* \xrightarrow{\mathbf{P}} 0$ in $\mathcal{D}_K$ for any compact rectangle $K \subset \mathbb{R}^3$ and that the sequence $(\tilde{E}_n^*)_{n=1}^\infty$ is tight in $\mathcal{D}_K$.

In what follows we will describe the limiting distribution of $\tilde{E}_n^*$, namely $\tilde{E}^*$, and show that the (unconditional) asymptotic distribution of $\tilde{\Delta}_n^*$ is that of the smallest argmax of $\tilde{E}^*$. This result will help us show that the ECDF bootstrap is inconsistent.

We start by introducing some notation. Recall the definitions of the random elements $\mathbf{Z}_1$, $\mathbf{Z}_2$, ν_1, ν_2, $\mathbf{u}$ and $\mathbf{v}$ as in the discussion preceding (5.11). Also let $\tau = (\tau_n)_{n=1}^\infty$ and $\kappa = (\kappa_n)_{n=1}^\infty$ two sequences of i.i.d. Poisson(1) random variables. Assume, in addition, that $\mathbf{Z}_1$, $\mathbf{Z}_2$, ν_1, ν_2, $\mathbf{v}$, $\mathbf{u}$, τ and κ are all mutually independent. Then, define the process $\tilde{\Xi} = (\tilde{\Xi}^{(1)}, \ldots, \tilde{\Xi}^{(6)})'$ as

$$\tilde{\Xi}(t) := \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \\ \sum_{0<j\leq \nu_1(-t)} \kappa_j \mathbf{1}_{t<0} \\ \sum_{0<j\leq \nu_1(-t)} \upsilon_j \kappa_j \mathbf{1}_{t<0} \\ \sum_{0<j\leq \nu_2(t)} \tau_j \mathbf{1}_{t\geq 0} \\ \sum_{0<j\leq \nu_2(t)} u_j \tau_j \mathbf{1}_{t\geq 0} \end{pmatrix} \tag{5.18}$$

for $t \in \mathbb{R}$ and let $\tilde{E}^*$ be given by

$$\begin{aligned} \tilde{E}^*(h) \quad = \quad & 2h_2\tilde{\Xi}^{(1)}(h_1) - h_2^2\mathbb{P}(Z \leq \zeta_0) + 2h_3\tilde{\Xi}^{(2)}(h_1) - h_3^2\mathbb{P}(Z > \zeta_0) \\ & + 2(\beta_0 - \alpha_0)\tilde{\Xi}^{(4)}(h_1) - (\alpha_0 - \beta_0)^2\tilde{\Xi}^{(3)}(h_1) \\ & + 2(\alpha_0 - \beta_0)\tilde{\Xi}^{(6)}(h_1) - (\alpha_0 - \beta_0)^2\tilde{\Xi}^{(5)}(h_1) \end{aligned} \tag{5.19}$$

for $h = (h_1, h_2, h_3) \in \mathbb{R}^3$. Additionally define the $\mathcal{S}$–valued (pure jump) processes $\tilde{J}_n$, $\tilde{J}_n^*$ and $\tilde{J}^*$ as

$$\tilde{J}_n^*(t) = \tilde{J}_n(t) \quad := \quad n\mathbb{P}_n^*(\mathbf{1}_{\zeta_0+\frac{t}{n}<Z\leq\zeta_0}) + n\mathbb{P}_n^*(\mathbf{1}_{\zeta_0<Z\leq\zeta_0+\frac{t}{n}}), \tag{5.20}$$

$$\tilde{J}^*(t) \quad := \quad \nu_1(-t)\mathbf{1}_{t<0} + \nu_2(t)\mathbf{1}_{t\geq 0}. \tag{5.21}$$

Lemma 5.4.6 (proved in Section 5.9.11) now states the asymptotic distribution of $\tilde{E}_n$ and of $n(\zeta_n^* - \zeta_0)$.

Lemma 5.4.6 *Consider the processes $\tilde{\Xi}_n$, $\tilde{E}_n$, $\tilde{J}_n$, $\tilde{\Xi}$, $\tilde{E}^*$ and $\tilde{J}^*$ as defined in (5.17), (5.16), (5.20), (5.18), (5.19) and (5.21), respectively. Then, unconditionally,*

(i) $\tilde{\Xi}_n \rightsquigarrow \tilde{\Xi}$ in $\mathbb{R}^2 \times \mathcal{D}_I^4$ for any compact interval $I \subset \mathbb{R}$;

(ii) $(\tilde{E}_n, \tilde{J}_n) \rightsquigarrow (\tilde{E}^*, \tilde{J}^*)$ in $\mathcal{D}_K \times \mathcal{S}_I$ for any compact interval $I \subset \mathbb{R}$ and any compact rectangle $K = A \times B \times I \subset \mathbb{R}^3$;

(iii) $\tilde{\Delta}_n^* = sargmax_{h\in\mathbb{R}^3}\{\tilde{E}_n(h)\} \rightsquigarrow sargmax_{h\in\mathbb{R}^3}\{\tilde{E}^*(h)\}$.

As a consequence, if the ECDF bootstrap is consistent, the variance of $sargmax_{h\in\mathbb{R}^3}\{\tilde{E}^*(h)\}$ *must be twice that of* $sargmax_{h\in\mathbb{R}^3}\{E^*(h)\}$.

As analytic expressions for the asymptotic variances of $n(\zeta_n^* - \zeta_0)$ and $n(\hat{\zeta}_n - \zeta_0)$ are not known, we use simulations to compute them. As an illustration, we take $\epsilon \sim N(0,1)$, $Z \sim N(0,1)$, $\alpha_0 = -1$, $\beta_0 = 1$ and $\zeta_0 = 0$ in (5.1). We approximate the limiting variances with the sample variances computed from 20,000 observations from each of the two asymptotic distributions. Our results are summarized in the following table, which immediately shows that the asymptotic variance of $n(\zeta_n^* - \zeta_0)$ is not twice that of $n(\hat{\zeta}_n - \zeta_0)$. Thus the ECDF bootstrap cannot be consistent.

Random variable	Asymptotic Variance
$n(\hat{\zeta}_n - \zeta_0)$	7.620948
$n(\zeta_n^* - \zeta_0)$	63.98377

5.4.2 Scheme 2 (Bootstrapping "residuals")

Another resampling procedure that arises naturally in a regression setup is bootstrapping "residuals". As with scheme 1, bootstrapping the "residuals" fixing the covariates is also *inconsistent*. Heuristically speaking, the resampling distribution fails to approximate the density of the predictor at the change-point ζ_0 at rate-n, and this leads to the inconsistency.

We recall the notation of Section 2. There we described the basic elements of the traditional fixed-design bootstrap of residuals and how to compute the bootstrap estimates θ_n^*. We first show that these bootstrap estimators converge conditionally in probability (almost surely) to the true value of the parameter. Then, we will provide a strong argument against the consistency of this bootstrap scheme. For notational convenience, we introduce the process R_n given by

$$R_n(\theta) := -\frac{1}{n}\sum_{j=1}^{n}\left(Y^*_{n,j} - \alpha \mathbf{1}_{Z_j\le\zeta} - \beta\mathbf{1}_{Z_j>\zeta}\right)^2 \quad \forall\, \theta \in \Theta.$$

We start by showing that the "centered" empirical distribution for the least squares residuals, $\mathbb{P}^{\epsilon}_n$, converges to the distribution of ϵ in total variation distance with probability one and its second moment is an almost surely consistent estimator of σ^2. This lemma will also be useful for the analysis of the smoothed bootstrap procedure. The proof can be found in Section 5.9.12.

Lemma 5.4.7 *Let G and φ be, respectively, the distribution and characteristic functions of ϵ. Then,*

(i) for any $\eta > 0$ we have that $\sup_{|\xi|\le\eta}\left\{\left|\int e^{i\xi x} d\mathbb{P}^{\epsilon}_n(x) - \varphi(\xi)\right|\right\} \xrightarrow{a.s.} 0;$

(ii) $\|\mathbb{P}^{\epsilon}_n - G\|_{\mathbb{R}} \xrightarrow{a.s.} 0$*;*

(iii) $\int x^2 d\mathbb{P}^{\epsilon}_n(x) \xrightarrow{a.s.} \sigma^2$*;*

(iv) $\int |x| d\mathbb{P}^{\epsilon}_n(x) \xrightarrow{a.s.} \mathbb{P}(|\epsilon|)$*;*

(v) if ϵ has a finite third moment under $\mathbb{P}$, then

$$\overline{\lim_{n\to\infty}} \int |x|^3 d\mathbb{P}^{\epsilon}_n(x) < \infty \quad \textit{almost surely.}$$

The next result (proved in Section 5.9.13) shows that the bootstrapped least squares estimators converge conditionally in probability with probability one.

Proposition 5.4.2 *Let $K \subset \Theta$ be a compact rectangle. Then,* *(i)* $\|R_n + \mathbb{P}_n^*(\tilde{\epsilon}_n^2) - M_n - \sigma^2\|_K \xrightarrow[a.s.]{\mathbf{P}_{\mathfrak{X}}} 0$;

(ii) $\|R_n + \mathbb{P}_n^*(\tilde{\epsilon}_n^2) - M - \sigma^2\|_K \xrightarrow[a.s.]{\mathbf{P}_{\mathfrak{X}}} 0$;

(iii) $\theta_n^* \xrightarrow[a.s.]{\mathbf{P}_{\mathfrak{X}}} \theta_0$ *and* $\theta_n^* - \hat{\theta}_n \xrightarrow[a.s.]{\mathbf{P}_{\mathfrak{X}}} 0$.

where M_n *and* M *are defined as in (5.3) and the subsequent paragraph.*

Consider the following process

$$\hat{E}_n(h) = -\sum_{j=1}^{n} \left(Y_{n,j}^* - \left(\hat{\alpha}_n + \frac{h_1}{\sqrt{n}} \right) \mathbf{1}_{Z_j \le \hat{\zeta}_n + \frac{h_3}{n}} - \left(\hat{\beta}_n + \frac{h_2}{\sqrt{n}} \right) \mathbf{1}_{Z_j > \hat{\zeta}_n + \frac{h_3}{n}} \right)^2 + \sum_{j=1}^{n} (\epsilon_{n,j}^*)^2.$$

Then for n large enough we have that

$$\left(n(\zeta_n^* - \hat{\zeta}_n), \sqrt{n}(\alpha_n^* - \hat{\alpha}_n), \sqrt{n}(\beta_n^* - \hat{\beta}_n) \right)' = \operatorname*{sargmax}_{h \in \mathbb{R}^3} \left\{ \hat{E}_n(h) \right\}.$$

Next we argue that the sequence $(\hat{E}_n)_{n=1}^{\infty}$ does not have a weak limit in probability and therefore distributional convergence of their corresponding smallest minimizers seems unreasonable. We refer the reader to Section 5.9.14 for a complete proof of the statement.

Lemma 5.4.8 *There is a compact rectangle* $K \subset \mathbb{R}^3$ *such that the sequence of processes* $(\hat{E}_n)_{n=1}^{\infty}$ *does not have a weak limit in probability in* $\mathcal{D}_K$.

5.5 Consistent bootstrap procedures

Here we will prove that the "smoothed bootstrap" (scheme 3) and the m out of n bootstrap (scheme 4) procedures yield consistent methods for constructing confidence intervals around the parameters.

5.5.1 Scheme 3 (Smoothed Bootstrap)

To show that scheme 3 (smoothed bootstrap + bootstrapping residuals) achieves consistency we appeal to Propositions 5.3.1, 5.3.2 and 5.3.3 by proving that the regularity conditions (I)-(VIII) of Section 3 hold for this scheme. Recall the description of this bootstrap procedure given in Section 5.2. Let $\hat{f}_n$ and $\hat{F}_n$ be the estimated smoothed density and distribution function of Z, respectively. For $I := [c, d] \subset \mathbb{R}$, a compact interval such that $\zeta_0 \in (c, d)$, we require the following two properties of $\hat{f}_n$ and $\hat{F}_n$:

$$\|\hat{F}_n - F\|_{\mathbb{R}} \xrightarrow{a.s.} 0; \tag{5.22}$$

$$\|\hat{f}_n - f\|_I \xrightarrow{a.s.} 0. \tag{5.23}$$

We would want to highlight that these conditions are fulfilled by many density estimation procedures. In particular, they hold when the density f is continuous and we let $\hat{f}_n$ be the kernel density estimator constructed from a suitable choice of kernel and bandwidth (e.g., see Silverman (1978)).

Let $\theta_n = \hat{\theta}_n$, $m_n = n$ and $\mathbb{Q}_n$ be the distribution that generates the bootstrap sample. Observe that under $\mathbb{Q}_n$, $\tilde{\epsilon}_n$ and Z are independent and that Z is a continuous random variable with density $\hat{f}_n$. The next result (proved in Section 5.9.15) shows that the bootstrapped least squares estimators achieve the right rate of convergence.

Proposition 5.5.1 *If (5.22) and (5.23) hold, then w.p.1, the sequence of conditional distributions of* $\left(n(\zeta_n^* - \hat{\zeta}_n), \sqrt{n}(\alpha_n^* - \hat{\alpha}_n), \sqrt{n}(\beta_n^* - \hat{\beta}_n)\right)'$ *is tight.*

Scheme 3 uses an approximation to the density of Z and this turns out to be crucial. The bootstrap measures now satisfy property (VI) on Section 5.3 and the bootstrap procedure is *strongly consistent*, as shown in the next result (proved in Section 5.9.16).

Proposition 5.5.2 *For scheme 3, provided that (5.22) and (5.23) hold, conditions (I)–(VIII) are satisfied with probability one, and thus,*

$$\begin{pmatrix} n(\zeta_n^* - \hat{\zeta}_n) \\ \sqrt{n}(\alpha_n^* - \hat{\alpha}_n) \\ \sqrt{n}(\beta_n^* - \hat{\beta}_n) \end{pmatrix} \rightsquigarrow \underset{h\in\mathbb{R}^3}{\mathit{sargmax}}\{E^*(h)\} \text{ almost surely.}$$

5.5.2 Scheme 4 (m out of n bootstrap)

For this scheme we will again use the framework established in Section 5.3. We take $(m_n)_{n=1}^{\infty}$ to be any sequence of natural numbers which increases to infinity, $\hat{\theta}_n = \theta_n$ and $\mathbb{Q}_n = \mathbb{P}_n$. The next result (proved in Section 5.9.17) shows the *weak consistency* of this procedure.

Proposition 5.5.3 *If $m_n = o(n)$ and $m_n \to \infty$, then conditions (I)–(VIII) hold (in probability) and we have*

$$\begin{pmatrix} n(\zeta_n^* - \hat{\zeta}_n) \\ \sqrt{m_n}(\alpha_n^* - \hat{\alpha}_n) \\ \sqrt{m_n}(\beta_n^* - \hat{\beta}_n) \end{pmatrix} \rightsquigarrow \underset{h\in\mathbb{R}^3}{\mathit{sargmax}}\{E^*(h)\} \text{ in probability.} \tag{5.24}$$

Remark: To prove Proposition 5.5.3, we will, in fact, show that for every subsequence $(n_k)_{k=1}^{\infty}$, there is a further subsequence $(n_{k_s})_{s=1}^{\infty}$, such that (I)-(VIII) hold w.p. 1 for $(n_{k_s})_{s=1}^{\infty}$ and (5.24) holds almost surely along the subsequence $(n_{k_s})_{s=1}^{\infty}$.

5.6 Simulation experiments

In this section we report the finite sample performance of the different bootstrap schemes on simulated data. We simulated random draws from four different models following (5.1). Each of these corresponded to choosing different pairs (F, G) of distributions for Z and ϵ (having mean 0), respectively. The pairs considered were $(N(0,2), N(0,1))$, $(4B(4,6)-2, N(0,1))$, $(4B(4,6)-2, \text{Unif}(-1,1))$, and $(4B(4,6)-2, \Gamma(4,2)-2)$, where $B(\cdot,\cdot)$ and $\Gamma(\cdot,\cdot)$ denote the beta and gamma distributions respectively.

For each of these models, we considered 1000 random samples of sizes $n = 50, 100, 200, 500$. For each sample, and for each of the bootstrap schemes, we took $4n$ bootstrap replicates to approximate the bootstrap distribution. The following table provides the estimated coverage proportions of nominal 95% CIs and average lengths of the CIs obtained using the 4 different bootstrap schemes for each of the four models.

At this point, we want to make some remarks about the computation of the estimators. We used a kernel density estimator based on the Gaussian kernel and chose the bandwidth by the so-called "normal reference rule" (see Scott (1992), page 131). In the case of the m out of n bootstrap, we did not use any data driven choice of m_n, but tried 3 different possibilities: $\lceil n^{\frac{4}{5}} \rceil, \lceil n^{\frac{9}{10}} \rceil$ and $\lceil n^{\frac{14}{15}} \rceil$. We will refer to the fixed-design bootstrapping of residuals scheme by FDR.

$Z \sim N(0,2), \epsilon \sim N(0,1)$						
Scheme	$n = 50$		$n = 200$		$n = 500$	
	Coverage	Avg Length	Coverage	Avg Length	Coverage	Avg Length
ECDF	0.83	1.14	0.79	0.22	0.81	0.08
Smoothed	0.94	0.94	0.95	0.19	0.95	0.07
FDR	0.83	0.76	0.86	0.16	0.90	0.06
$\lceil n^{4/5} \rceil$	0.87	0.87	0.91	0.23	0.91	0.08
$\lceil n^{9/10} \rceil$	0.85	1.02	0.87	0.21	0.87	0.079
$\lceil n^{14/15} \rceil$	0.85	1.05	0.84	0.21	0.86	0.08

$Z \sim 4B(4,6)-2, \epsilon \sim N(0,1)$						
Scheme	$n=50$		$n=200$		$n=500$	
	Coverage	Avg Length	Coverage	Avg Length	Coverage	Avg Length
ECDF	0.80	0.54	0.80	0.11	0.81	0.04
Smoothed	0.96	0.46	0.94	0.11	0.95	0.47
FDR	0.73	0.32	0.77	0.08	0.79	0.03
$\lceil n^{4/5} \rceil$	0.88	0.53	0.89	0.11	0.90	0.04
$\lceil n^{9/10} \rceil$	0.85	0.54	0.86	0.11	0.88	0.04
$\lceil n^{14/15} \rceil$	0.83	0.55	0.84	0.11	0.87	0.04
$Z \sim 4B(4,6)-2, \epsilon \sim \mathrm{Unif}(-1,1)$						
Scheme	$n=50$		$n=200$		$n=500$	
	Coverage	Avg Length	Coverage	Avg Length	Coverage	Avg Length
ECDF	0.80	0.40	0.80	0.08	0.81	0.03
Smoothed	0.94	0.33	0.95	0.08	0.96	0.04
FDR	0.75	0.26	0.77	0.06	0.81	0.02
$\lceil n^{4/5} \rceil$	0.88	0.36	0.88	0.09	0.91	0.04
$\lceil n^{9/10} \rceil$	0.85	0.39	0.85	0.08	0.87	0.03
$\lceil n^{14/15} \rceil$	0.83	0.39	0.84	0.08	0.85	0.03
$Z \sim 4B(4,6)-2, \epsilon \sim \Gamma(4,2)-2$						
Scheme	$n=50$		$n=200$		$n=500$	
	Coverage	Avg Length	Coverage	Avg Length	Coverage	Avg Length
ECDF	0.80	0.49	0.80	0.09	0.81	0.04
Smoothed	0.93	0.36	0.95	0.08	0.96	0.03
FDR	0.76	0.30	0.77	0.06	0.80	0.02
$\lceil n^{4/5} \rceil$	0.87	0.43	0.88	0.10	0.91	0.03
$\lceil n^{9/10} \rceil$	0.85	0.46	0.84	0.09	0.88	0.03
$\lceil n^{14/15} \rceil$	0.83	0.48	0.85	0.09	0.85	0.03

We can see from the table that the smoothed bootstrap scheme outperforms all the others in terms of coverage. It must also be noted that this is achieved without a relative increase in the lengths of the intervals. The m out of n bootstrap with $\lceil n^{4/5} \rceil$ also performs reasonably well. It clearly outperforms all other m out of n schemes as well as ECDF and FDR bootstrap procedures (which are inconsistent).

Figure 5.1 shows the histograms of the distribution of $n(\hat{\zeta}_n - \zeta_0)$ (obtained from 1000 random samples) and its bootstrap estimates obtained from the 4 different bootstrap schemes (using 2000 bootstrap samples each) from a single data set of size $n = 500$ from model (5.1) with $Z \sim 4B(4,6) - 2, \epsilon \sim \Gamma(4,2) - 2, \alpha_0 = -1, \beta_0 = 1, \zeta_0 = 0$. The histograms clearly show that the smoothed bootstrap (top right panel) provides, by far, the best approximation to both, the actual (top middle panel) and the limiting distributions (top left panel). In fact, the histograms of the distribution of $n(\hat{\zeta}_n - \zeta_0)$ and the corresponding smoothed bootstrap estimate are almost indistinguishable. The m out of n approach, although guaranteed to converge, lacks the efficiency of the smoothed bootstrap. This may be due to the fact that we do not have an optimal way of choosing the tuning parameter m_n. The smoothed bootstrap also requires the choice of a tuning parameter, namely, the smoothing bandwidth, but the in our analysis the results were very insensitive to the choice of the bandwidth. This is certainly an advantage for the smoothed bootstrap procedure.

5.7 More general change-point regression models

In this section we mention some of the broader implications of our analysis of (5.1) in the context of more general change-point models in regression. We can consider a model of the form

$$Y = \psi_{\alpha_0}(W, Z)\mathbf{1}_{Z \le \zeta_0} + \xi_{\beta_0}(W, Z)\mathbf{1}_{Z > \zeta_0} + \epsilon, \tag{5.25}$$

where Z is a continuous random variable; W is a random vector of covariates; $\alpha_0 \in \mathbb{R}^p$ and $\beta_0 \in \mathbb{R}^q$ are two unknown Euclidian parameters; $\psi_\alpha(w, z)$ and $\xi_\beta(w, z)$ are known real-valued functions continuous in (w, z) and twice continuously differentiable in α and β respectively; $\zeta_0 \in [a, b] \subset \text{supp}(Z) \subset \mathbb{R}$ is the change-point; ϵ is a continuous random variable, independent of (W, Z)

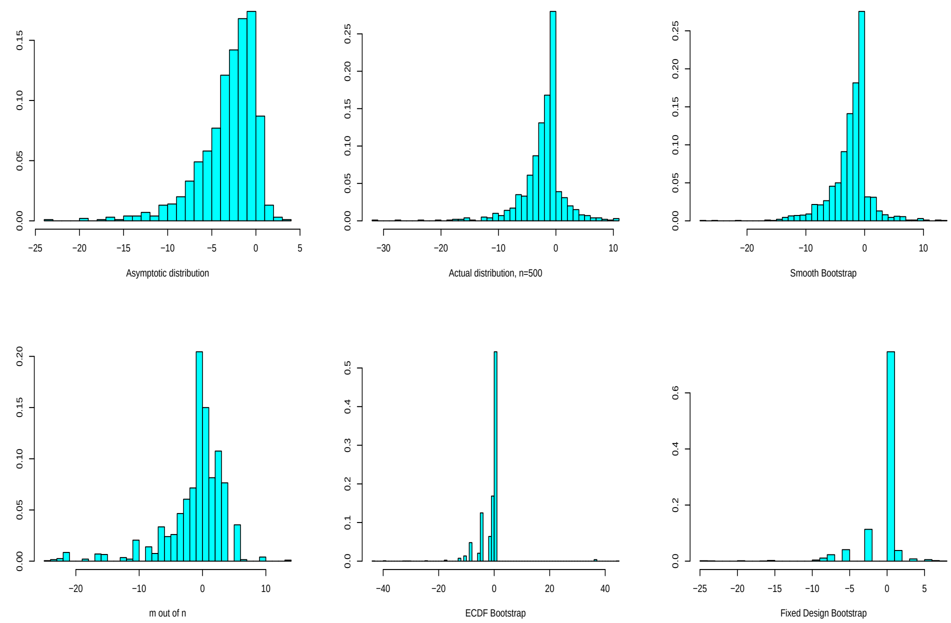

Figure 5.1: Histograms of the distribution of $n(\hat{\zeta}_n - \zeta_0)$ and its bootstrap estimates: the asymptotic distribution of $n(\hat{\zeta}_n - \zeta_0)$ (top left); the actual distribution of $n(\hat{\zeta}_n - \zeta_0)$ (top middle); the distribution of $n(\zeta_n^* - \hat{\zeta}_n)$ for the smoothed (top right), ECDF (bottom middle) and FDR (bottom right) schemes; the distribution of $m_n(\zeta_n^* - \hat{\zeta}_n)$, $m_n = \lceil n^{\frac{4}{5}} \rceil$ (bottom left).

with zero expectation and finite variance $\sigma^2 > 0$. We assume that $\psi_{\alpha_0}(W, Z)$ is identifiable from $\xi_{\beta_0}(W, Z)$ and that the least squares problems

$$\min_{\alpha \in \mathbb{R}^p} \left\{ \sum_{Z_j \le \zeta} (Y_j - \psi_\alpha(W_j, Z_j))^2 \right\} \quad \text{and} \quad \min_{\beta \in \mathbb{R}^q} \left\{ \sum_{Z_j > \zeta} (Y_j - \xi_\beta(W_j, Z_j))^2 \right\}$$

are well-posed for every possible data set $\{(Y_1, Z_1, W_1), \ldots, (Y_n, Z_n, W_n)\}$ and any $\zeta \in \mathrm{supp}(Z)^\circ$. We also assume that $\psi_{\alpha_0}(w, \zeta_0) \neq \xi_{\beta_0}(w, \zeta_0)$ for every value of w.

Like in the simple case, the method of least squares can be used to compute estimators $\hat{\alpha}_n$, $\hat{\beta}_n$ and $\hat{\zeta}_n$. One simply takes the minimizer $(\hat{\alpha}_n, \hat{\beta}_n, \hat{\zeta}_n)$

of

$$\sum_{j=1}^{n} \left(Y_j - \psi_\alpha(W_j, Z_j)\mathbf{1}_{Z_j \le \zeta} + \xi_\beta(W_j, Z_j)\mathbf{1}_{Z_j > \zeta}\right)^2$$

with the smallest ζ-component.

Since the simple model (5.1) is a particular case of (5.25), one can immediately conclude from our analysis that the usual ECDF and residual bootstrap procedures will not be consistent. However, the smoothed bootstrap can be adapted to produce consistent interval estimation. The modified scheme can be described as follows:

1. Choose some procedure (e.g., kernel density estimation) to build a distribution $\hat{F}_n$ with density $\hat{f}_n$ such that $\hat{f}_n \to f$ uniformly on some open interval containing ζ_0 w.p. 1, where f is the density of Z. Let $\mathbb{P}_n^\epsilon$ and $\mathbb{P}_n^W$ be the empirical measures of the centered residuals (as in the description of Scheme 2 in Section 5.2) and $W_1, \ldots, W_n$, respectively.

2. Get i.i.d. replicates $Z_{n,1}^*, \ldots, Z_{n,n}^*$ from $\hat{F}_n$ and sample, independently, $\epsilon_{n,1}^*, \ldots, \epsilon_{n,n}^* \overset{i.i.d.}{\sim} \mathbb{P}_n^\epsilon$ and $W_{n,1}^*, \ldots, W_{n,n}^* \overset{i.i.d.}{\sim} \mathbb{P}_n^W$. Here we can also keep W_i's fixed, i.e., $W_{n,i}^* = W_i$.

3. Define $Y_{n,j}^* = \psi_{\hat{\alpha}_n}(W_{n,j}^*, Z_{n,j}^*)\mathbf{1}_{Z_{n,j}^* \le \hat{\zeta}_n} + \xi_{\hat{\beta}_n}(W_{n,j}^*, Z_{n,j}^*)\mathbf{1}_{Z_{n,j}^* > \hat{\zeta}_n} + \epsilon_{n,j}^*$ for all $j = 1, \ldots, n$.

4. Compute the bootstrap least squares estimators $(\alpha_n^*, \beta_n^*, \zeta_n^*)$ by taking the minimizer of

$$\sum_{j=1}^{n} \left(Y_{n,j}^* - \psi_\alpha(W_{n,j}^*, Z_{n,j}^*)\mathbf{1}_{Z_{n,j}^* \le \zeta} - \xi_\beta(W_{n,j}^*, Z_{n,j}^*)\mathbf{1}_{Z_{n,j}^* > \zeta}\right)^2$$

   with the smallest ζ-component.

5. Approximate the distribution of $n(\hat{\zeta}_n - \zeta_0)$ with the (conditional) distribution of $n(\zeta_n^* - \hat{\zeta}_n)$.

Although our analysis indicates that this smoothed bootstrap procedure must be consistent, it is difficult to use our methods to prove consistency in such generality. However, the proof of consistency for the simple model (5.1) can be adapted to cover the case of parametric additive models, i.e., when $\psi_\alpha(w,z)$ and $\xi_\beta(w,z)$ are of the form

$$\psi_\alpha(w,z) = \sum_{j=1}^{p} \alpha_j g_j(w,z), \quad \text{and} \quad \xi_\beta(w,z) = \sum_{k=1}^{q} \beta_k h_k(w,z),$$

where g_j, h_k, $j = 1,\ldots,p$, $k = 1,\ldots,q$ are known smooth functions.

5.8 Acknowledgement

We would like to thank Souvik Ghosh for his helpful comments on the proof of Lemma 5.4.6.

5.9 Supplementary Lemmas

In this section we provide the proofs of most of the results stated in the previous sections. We start by giving an account of a series of technical lemmas which will aid us in the proof of Propositions 5.3.1, 5.3.2 and 5.3.3.

Lemma 5.9.1 *Let $\alpha \neq \beta \in \mathbb{R}$. Consider the class of functions from $\mathbb{R}^2$ to $\mathbb{R}$ given by*

$$\mathscr{A} = \left\{\phi(y,z) := (y - \alpha \mathbf{1}_{(-\infty,\zeta]}(z) - \beta \mathbf{1}_{(\zeta,\infty]}(z))\mathbf{1}_I(z) | \zeta \in \mathbb{R}, I \subset \mathbb{R} \textit{ is an interval}\right\}.$$

Then, $\mathscr{A}$ is a VC-subgraph class with envelope $|y| + |\alpha| + |\beta|$. There is an upper bound for the VC-index of $\mathscr{A}$ that is independent of α and β. Moreover, there is a continuous, increasing function $J_{\mathscr{A}}$ with $J_{\mathscr{A}}(1) < \infty$, which is also independent of α and β, and satisfies the following property: If $\mathscr{D} \subset \mathscr{A}$ is *a subclass with envelope B and $W_1, \ldots, W_n$ is a random sample, defined on some probability space $(\Omega, \mathcal{A}, \mathbf{P})$, from a distribution μ for which $\mu(B^2) < \infty$ and μ_n is the empirical measure defined by the sample, then*

$$\int \sup_{\varphi \in \mathcal{D}} \{|(\mu_n - \mu)(\varphi)|\} \, d\mathbf{P} \leq \frac{J_{\mathscr{A}}(1)}{\sqrt{n}} \sqrt{\mu(B^2)}.$$

Proof: We use the same notation as in Lemmas 2.6.17 and 2.6.18, page 147 of van der Vaart and Wellner (1996). Consider the classes of functions $\mathscr{H} = \{y - \alpha \mathbf{1}_{(-\infty,\zeta]}(z) - \beta \mathbf{1}_{(\zeta,\infty]}(z) : \zeta \in \mathbb{R}\}$ and $\mathscr{K} = \left\{\mathbf{1}_{(-\infty,\zeta]}(z) : \zeta \in \mathbb{R}\right\}$. Then, $\mathscr{K}$ is a VC class with VC-index 2. It follows that $\mathscr{H} = (\beta - \alpha) \cdot \mathscr{K} + (y - \beta)$ is also VC. Recall that $\mathcal{F} = \{\mathbf{1}_I(z) : I \subset \mathbb{R} \text{ is an interval}\}$. Letting $[\varphi > t] := \{(y, z, t) : \varphi(y, z) > t\}$ for $\varphi \in \mathscr{A}$, we see that

$$\{[\varphi > t] : \varphi \in \mathscr{A}\} = \Big(\mathbb{R} \times \{\mathcal{F} \leq 0\} \times (-\infty, 0)\Big) \bigsqcup \Big(\{[\psi > t] : \psi \in \mathscr{H}\} \sqcap (\mathbb{R} \times \{\mathcal{F} > 0\} \times \mathbb{R})\Big)$$

from which it follows that $\mathscr{A}$ is VC. Moreover, the VC-indexes of $\mathscr{K}$ and $\mathcal{F}$ are two and three for any choice of α and β. Hence, the corresponding VC-indexes of $\mathscr{H}$ and $\mathscr{A}$ both have upper bounds independent of α and β. The existence of the function $J_{\mathscr{A}}$ is a consequence of the maximal inequality 3.1 in Kim and Pollard (1990). Note that $J_{\mathscr{A}}$ only depends on the VC-index of the class $\mathscr{A}$, which in turn has an upper bound independent of α and β. $\square$

Lemma 5.9.2 *Suppose that (I)-(IV) hold. Then,*

(i) $\left\|\mathbb{Q}_n(\tilde{\epsilon}_n^2 \mathbf{1}_{Z \leq (\cdot) \wedge \zeta_n}) - \sigma^2 \mathbb{P}(Z \leq (\cdot) \wedge \zeta_0)\right\|_{[a,b]} \to 0,$

(ii) $\left\|\mathbb{Q}_n(|\tilde{\epsilon}_n| \mathbf{1}_{Z \leq (\cdot) \wedge \zeta_n}) - \mathbb{P}(|\epsilon|)\mathbb{P}(Z \leq (\cdot) \wedge \zeta_0)\right\|_{[a,b]} \to 0,$

(iii) $\left\|\mathbb{Q}_n(\tilde{\epsilon}_n \mathbf{1}_{Z \leq (\cdot) \wedge \zeta_n})\right\|_{[a,b]} \to 0,$ *and*

(iv) $\left\|\mathbb{Q}_n(\mathbf{1}_{Z\le(\cdot)\wedge\zeta_n}) - \mathbb{P}(\mathbf{1}_{Z\le(\cdot)\wedge\zeta_0})\right\|_{[a,b]} \to 0.$

Also, these statements are true if $\mathbf{1}_{Z\le(\cdot)\wedge\zeta_n}$ *is replaced by any of* $\mathbf{1}_{(\cdot)<Z\le\zeta_n}$*,* $\mathbf{1}_{\zeta_n<Z\le(\cdot)}$ *or* $\mathbf{1}_{Z>(\cdot)\vee\zeta_n}$*.*

Proof: Since $\zeta_n \to \zeta_0$ and Z is continuous, for any $\zeta \in [a,b]$, we obtain

$$\begin{aligned}
\left|\mathbb{P}(Y^2\mathbf{1}_{Z\le\zeta\wedge\zeta_n}) - \mathbb{P}(Y^2\mathbf{1}_{Z\le\zeta\wedge\zeta_0})\right| &\le \mathbb{P}(Y^2|\mathbf{1}_{Z\le\zeta_n} - \mathbf{1}_{Z\le\zeta_0}|) \to 0,\\
|\mathbb{P}(|Y-\alpha_0|\mathbf{1}_{Z\le\zeta\wedge\zeta_n}) - \mathbb{P}(|Y-\alpha_0|\mathbf{1}_{Z\le\zeta\wedge\zeta_0})| &\le \mathbb{P}(|Y||\mathbf{1}_{Z\le\zeta_n} - \mathbf{1}_{Z\le\zeta_0}|) \to 0,\\
|\mathbb{P}(Y\mathbf{1}_{Z\le\zeta\wedge\zeta_n}) - \mathbb{P}(Y\mathbf{1}_{Z\le\zeta\wedge\zeta_0})| &\le \mathbb{P}(|Y||\mathbf{1}_{Z\le\zeta_n} - \mathbf{1}_{Z\le\zeta_0}|) \to 0,\\
|\mathbb{P}(\mathbf{1}_{Z\le\zeta\wedge\zeta_n}) - \mathbb{P}(\mathbf{1}_{Z\le\zeta\wedge\zeta_0})| &\le \mathbb{P}(|\mathbf{1}_{Z\le\zeta_n} - \mathbf{1}_{Z\le\zeta_0}|) \to 0.
\end{aligned}$$

Also note that the convergence is uniform in $\zeta \in [a,b]$. Thus,

$$\begin{aligned}
&\left\|\mathbb{Q}_n(Y^2\mathbf{1}_{Z\le(\cdot)\wedge\zeta_n}) - \mathbb{P}(Y^2\mathbf{1}_{Z\le(\cdot)\wedge\zeta_0})\right\|_{[a,b]} \le \|\mathbb{Q}_n - \mathbb{P}\|_{\mathcal{H}}\\
&\qquad + \left\|\mathbb{P}(Y^2\mathbf{1}_{Z\le(\cdot)\wedge\zeta_n}) - \mathbb{P}(Y^2\mathbf{1}_{Z\le(\cdot)\wedge\zeta_0})\right\|_{[a,b]} \to 0
\end{aligned}$$

as $n \to \infty$ by (III). Similarly, we also obtain that $\|\mathbb{Q}_n(|Y-\alpha_0|\mathbf{1}_{Z\le(\cdot)\wedge\zeta_n}) - \mathbb{P}(|Y-\alpha_0|\mathbf{1}_{Z\le(\cdot)\wedge\zeta_0})\|_{[a,b]} \to 0$, $\|\mathbb{Q}_n(Y\mathbf{1}_{Z\le(\cdot)\wedge\zeta_n}) - \mathbb{P}(Y\mathbf{1}_{Z\le(\cdot)\wedge\zeta_0})\|_{[a,b]} \to 0$ and $\|\mathbb{Q}_n(\mathbf{1}_{Z\le(\cdot)\wedge\zeta_n}) - \mathbb{P}(\mathbf{1}_{Z\le(\cdot)\wedge\zeta_0})\|_{[a,b]} \to 0$. This proves (iv).

Finally, (i), (ii) and (iii) follow as consequence of the convergence $\alpha_n \to \alpha_0$ and of the following inequalities:

$$\begin{aligned}
&\left\|\mathbb{Q}_n(\tilde{\epsilon}_n^2\mathbf{1}_{Z\le(\cdot)\wedge\zeta_n}) - \sigma^2\mathbb{P}(Z\le(\cdot)\wedge\zeta_0)\right\|_{[a,b]}\\
\le\ &\left\|\mathbb{Q}_n(Y^2\mathbf{1}_{Z\le(\cdot)\wedge\zeta_n}) - \mathbb{P}(Y^2\mathbf{1}_{Z\le(\cdot)\wedge\zeta_0})\right\|_{[a,b]} + 2|\alpha_n-\alpha_0|\mathbb{Q}_n(|Y|) + |\alpha_n^2 - \alpha_0^2|\\
+\ &2|\alpha_0|\left\|\mathbb{Q}_n(Y\mathbf{1}_{Z\le(\cdot)\wedge\zeta_n}) - \mathbb{P}(Y\mathbf{1}_{Z\le(\cdot)\wedge\zeta_0})\right\|_{[a,b]} + \alpha_0^2\left\|\mathbb{Q}_n(\mathbf{1}_{Z\le(\cdot)\wedge\hat{\zeta}_n}) - \mathbb{P}(\mathbf{1}_{Z\le(\cdot)\wedge\zeta_0})\right\|_{[a,b]}
\end{aligned}$$

and

$$\begin{aligned}
\left\|\mathbb{Q}_n(|\tilde{\epsilon}_n|\mathbf{1}_{Z\le(\cdot)\wedge\zeta_n}) - \mathbb{P}(|\epsilon|\mathbf{1}_{Z\le(\cdot)\wedge\zeta_n})\right\|_{[a,b]} &\le\\
\left\|\mathbb{Q}_n(|Y-\alpha_0|\mathbf{1}_{Z\le(\cdot)\wedge\zeta_n}) - \mathbb{P}(|Y-\alpha_0|\mathbf{1}_{Z\le(\cdot)\wedge\zeta_0})\right\|_{[a,b]} &+ |\alpha_n - \alpha_0|.
\end{aligned}$$

and

$$\left\|\mathbb{Q}_n(\tilde{\epsilon}_n \mathbf{1}_{Z\leq(\cdot)\wedge\zeta_n})\right\|_{[a,b]} \leq \left\|\mathbb{Q}_n(Y\mathbf{1}_{Z\leq(\cdot)\wedge\zeta_n}) - \mathbb{P}(Y\mathbf{1}_{Z\leq(\cdot)\wedge\zeta_0})\right\|_{[a,b]}$$
$$+|\alpha_n - \alpha_0| + |\alpha_0| \left\|\mathbb{Q}_n(\mathbf{1}_{Z\leq(\cdot)\wedge\zeta_n}) - \mathbb{P}(\mathbf{1}_{Z\leq(\cdot)\wedge\zeta_0})\right\|_{[a,b]}.$$

The other three cases follow from similar arguments. □

Lemma 5.9.3 *Suppose that (I)-(IV) hold. Then,*

(i) $\left\|(\mathbb{P}_n^* - \mathbb{Q}_n)(\tilde{\epsilon}_n \mathbf{1}_{Z\leq(\cdot)\wedge\zeta_n})\right\|_{[a,b]} \xrightarrow{\mathbf{P}} 0,$

(ii) $\left\|(\mathbb{P}_n^* - \mathbb{Q}_n)(\mathbf{1}_{Z\leq(\cdot)\wedge\zeta_n})\right\|_{[a,b]} \xrightarrow{\mathbf{P}} 0.$

Also, these statements are true if $\mathbf{1}_{Z\leq(\cdot)\wedge\zeta_n}$ *is replaced by any of* $\mathbf{1}_{(\cdot)<Z\leq\zeta_n}$, $\mathbf{1}_{\zeta_n<Z\leq(\cdot)}$ *or* $\mathbf{1}_{Z>(\cdot)\vee\zeta_n}$.

Proof: By the maximal inequality 3.1 from Kim and Pollard (1990) and Lemma 5.9.1 we see that:

$$\begin{aligned}\mathbf{E}\left(\left\|(\mathbb{P}_n^* - \mathbb{Q}_n)(\tilde{\epsilon}_n \mathbf{1}_{Z\leq(\cdot)\wedge\zeta_n})\right\|_{[a,b]}\right) &\leq \frac{J_{\mathscr{A}}(1)}{\sqrt{m_n}}\sqrt{\mathbb{Q}_n(\tilde{\epsilon}_n^2)}\\ \mathbf{E}\left(\left\|(\mathbb{P}_n^* - \mathbb{Q}_n)(\mathbf{1}_{Z\leq(\cdot)\wedge\zeta_n})\right\|_{[a,b]}\right) &\leq \frac{J_{\mathcal{F}}(1)}{\sqrt{m_n}}.\end{aligned}$$

The lemma now follow directly as $\mathbb{Q}_n(\tilde{\epsilon}_n^2) \to \sigma^2$ (a consequence of Lemma 5.9.2). The other statements are proven similarly. □

5.9.1 Proof of Proposition 5.3.1

Noting that $\tilde{\epsilon}_n = Y - \alpha_n \mathbf{1}_{Z\leq\zeta_n} - \beta_n \mathbf{1}_{Z>\zeta_n}$, we write

$$\begin{aligned}m_\theta(X) &= -(\tilde{\epsilon}_n + \alpha_n - \alpha)^2 \mathbf{1}_{Z\leq\zeta_n\wedge\zeta} - (\tilde{\epsilon}_n + \beta_n - \alpha)^2 \mathbf{1}_{\zeta_n<Z\leq\zeta}\\ &\quad -(\tilde{\epsilon}_n + \alpha_n - \beta)^2 \mathbf{1}_{\zeta<Z\leq\zeta_n} - (\tilde{\epsilon}_n + \beta_n - \beta)^2 \mathbf{1}_{Z>\zeta_n\vee\zeta},\end{aligned} \tag{5.26}$$

and therefore

$$\begin{aligned} -\mathbb{P}_n^*(\tilde{\epsilon}_n^2) &= M_n^*(\theta_n) \qquad \leq M_n^*(\theta_n^*) \\ &\leq -\mathbb{P}_n^*[(\tilde{\epsilon}_n - \alpha_n^* + \alpha_n)^2 \mathbf{1}_{Z<a}] - \mathbb{P}_n^*[(\tilde{\epsilon}_n - \beta_n^* + \beta_n)^2 \mathbf{1}_{Z>b}]. \end{aligned}$$

Letting $\gamma_n^* = (\alpha_n^*, \beta_n^*)$, noticing that $M_n^*(\hat{\theta}_n) = -\mathbb{P}_n^*(\tilde{\epsilon}_n^2)$, and by rearranging the terms in the above inequality, we get

$$\begin{aligned} &|\gamma_n^* - \gamma_n|^2 \mathbb{P}_n^*(Z<a) \wedge \mathbb{P}_n^*(Z>b) \leq \mathbb{P}_n^*\left(\tilde{\epsilon}_n^2 \mathbf{1}_{a\leq Z\leq b}\right) \\ &\qquad +2|\gamma_n^* - \gamma_n| \left(|\mathbb{P}_n^*\left(\tilde{\epsilon}_n \mathbf{1}_{Z<a}\right)| + |\mathbb{P}_n^*\left(\tilde{\epsilon}_n \mathbf{1}_{Z>b}\right)|\right). \end{aligned}$$

Consider $\mathbb{P}_n^*(Z<a)$. By (ii) of Lemma 5.9.3 we see that $|(\mathbb{P}_n^* - \mathbb{Q}_n)(Z < a)| \xrightarrow{\mathbf{P}} 0$ and by (iv) of Lemma 5.9.2 we can show that $|(\mathbb{Q}_n - \mathbb{P})(Z<a)| \to 0$. Thus, combining the two, we have $\mathbb{P}_n^*(Z<a) \xrightarrow{\mathbf{P}} \mathbb{P}(Z<a)$. Similarly, we can show that $\mathbb{P}_n^*(Z<a) \wedge \mathbb{P}_n^*(Z>b) \xrightarrow{\mathbf{P}} \mathbb{P}(Z<a) \wedge \mathbb{P}(Z>b) > 0$ and also that $|\mathbb{P}_n^*\left(\tilde{\epsilon}_n \mathbf{1}_{Z<a}\right)| + |\mathbb{P}_n^*\left(\tilde{\epsilon}_n \mathbf{1}_{Z>b}\right)| \xrightarrow{\mathbf{P}} 0$. Also, observe that $\mathbf{E}\left(\mathbb{P}_n^*(\tilde{\epsilon}_n^2)\right) = \mathbb{Q}_n(\tilde{\epsilon}_n^2) \to \sigma^2$, by assumptions (I)-(III) and so $\mathbb{P}_n^*(\tilde{\epsilon}_n^2)$ is bounded in $\mathbb{L}^1$. Hence, we can write

$$|\gamma_n^* - \gamma_n|^2 \leq O_{\mathbf{P}}(1) + |\gamma_n^* - \gamma_n| o_{\mathbf{P}}(1)$$

and therefore $|\gamma_n^* - \gamma_n| = O_{\mathbf{P}}(1)$ (and, consequently, $|\gamma_n^* - \gamma_0| = O_{\mathbf{P}}(1)$).

We first rewrite $m_\theta(X)$ as follows:

$$\begin{aligned} m_\theta(X) = -\tilde{\epsilon}_n^2 &- 2(\alpha_n - \alpha)\tilde{\epsilon}_n \mathbf{1}_{Z\leq\zeta\wedge\zeta_n} - (\alpha_n - \alpha)^2 \mathbf{1}_{Z\leq\zeta\wedge\zeta_n} \\ &-2(\beta_n - \alpha)\tilde{\epsilon}_n \mathbf{1}_{\zeta_n<Z\leq\zeta} - (\beta_n - \alpha)^2 \mathbf{1}_{\zeta_n<Z\leq\zeta} \\ &-2(\alpha_n - \beta)\tilde{\epsilon}_n \mathbf{1}_{\zeta<Z\leq\zeta_n} - (\alpha_n - \beta)^2 \mathbf{1}_{\zeta<Z\leq\zeta_n} \\ &-2(\beta_n - \beta)\tilde{\epsilon}_n \mathbf{1}_{Z>\zeta\vee\zeta_n} - (\beta_n - \beta)^2 \mathbf{1}_{Z>\zeta\vee\zeta_n}. \end{aligned} \tag{5.27}$$

We can then decompose M_n^* as in (5.27), and use Lemmas 5.9.3 and 5.9.2 and the fact that $\theta_n \to \theta_0$, to obtain

$$\begin{aligned} \left\| M_n^* + \mathbb{P}_n^*(\tilde{\epsilon}_n^2) - M_n - \mathbb{Q}_n(\tilde{\epsilon}_n^2) \right\|_K &\xrightarrow{\mathbf{P}} 0. \\ \left\| M_n^* + \mathbb{P}_n^*(\tilde{\epsilon}_n^2) - M - \sigma^2 \right\|_K &\xrightarrow{\mathbf{P}} 0 \end{aligned}$$

for every compact $K \subset \Theta$. But θ_0 is also the unique maximizer of $M + \sigma^2$ and $|\gamma_n^* - \gamma_0| = O_{\mathbf{P}}(1)$. Therefore, the conditions of Corollary 3.2.3 (ii), page 287 of van der Vaart and Wellner (1996), hold and we obtain that $\theta_n^* \xrightarrow{\mathbf{P}} \theta_0$ (and also that $\theta_n^* - \theta_n \xrightarrow{\mathbf{P}} 0$). $\square$

5.9.2 Proof of Proposition 5.3.2

We will apply Theorem 3.4.1 of van der Vaart and Wellner (1996) to prove the result. Let $d : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$ be given by $d(\theta, \vartheta) = |(\theta_2, \theta_3) - (\vartheta_2, \vartheta_3)| + \sqrt{|\theta_1 - \vartheta_1|}$. Consider $\eta, \rho, L > 0$ as in (V) and a compact rectangle $K \subset \Theta$ such that $\{\theta \in \Theta : \ d(\theta, \theta_n) < \eta \text{ for some } n \in \mathbb{N}\} \subset K$. We can take L large enough so $L > 1 \vee \sup\{|\theta_3 - \vartheta_2| \vee |\theta_2 - \vartheta_3| : \theta, \vartheta \in K\}$. Pick n large enough so we can fix some $\delta \in (\frac{2\sqrt{2}}{m_n^{1/4}}, \eta)$. Then, taking also (I)-(IV) into account and possibly making η smaller, we can find positive constants $c_1, c_2 > 0$ and $N \in \mathbb{N}$ such that for any $n \geq N$, we have (5.5), (5.6), (5.7) and the inequalities:

$$\begin{aligned} \inf_{d(\theta,\theta_n)<\delta} \left\{|\alpha_n - \beta|^2 \wedge |\beta_n - \alpha|^2\right\} &> c_1, \\ \mathbb{Q}_n(Z \leq a) \wedge \mathbb{Q}_n(Z > b) &> c_2. \end{aligned}$$

Also, let $\mathbb{M}_n(\theta) := M_n^*(\theta) + \mathbb{P}_n^*(\tilde{\epsilon}_n^2)$ and $\mathcal{M}_n(\theta) := M_n(\theta) + \mathbb{Q}_n(\tilde{\epsilon}_n^2)$ for all $\theta \in \Theta$.

Choose $n \geq N$ and $\theta \in \Theta$ with $\frac{\delta}{2} < d(\theta, \theta_n) < \delta$. Then, considering the properties of the constants just defined and the expression

$$\begin{aligned} \mathcal{M}_n(\theta) - \mathcal{M}_n(\theta_n) = & -2(\alpha_n - \alpha)\mathbb{Q}_n(\tilde{\epsilon}_n \mathbf{1}_{Z \leq \zeta \wedge \zeta_n}) - (\alpha_n - \alpha)^2 \mathbb{Q}_n(\mathbf{1}_{Z \leq \zeta \wedge \zeta_n}) \\ & - 2(\beta_n - \alpha)\mathbb{Q}_n(\tilde{\epsilon}_n \mathbf{1}_{\zeta_n < Z \leq \zeta}) - (\beta_n - \alpha)^2 \mathbb{Q}_n(\mathbf{1}_{\zeta_n < Z \leq \zeta}) \\ & - 2(\alpha_n - \beta)\mathbb{Q}_n(\tilde{\epsilon}_n \mathbf{1}_{\zeta < Z \leq \zeta_n}) - (\alpha_n - \beta)^2 \mathbb{Q}_n(\mathbf{1}_{\zeta < Z \leq \zeta_n}) \\ & - 2(\beta_n - \beta)\mathbb{Q}_n(\tilde{\epsilon}_n \mathbf{1}_{Z > \zeta \vee \zeta_n}) - (\beta_n - \beta)^2 \mathbb{Q}_n(\mathbf{1}_{Z > \zeta \vee \zeta_n}) \end{aligned} \tag{5.28}$$

it is seen that the sum of the 1st, 3rd, 5th, and 7th terms in (5.28) can be bounded from above by $\frac{8L^2\delta}{\sqrt{m_n}}$. While we also have,

$$
\begin{aligned}
(\alpha_n-\alpha)^2\mathbb{Q}_n(\mathbf{1}_{Z\le\zeta\wedge\zeta_n}) &\ge c_2(\alpha_n-\alpha)^2,\\
(\beta_n-\beta)^2\mathbb{Q}_n(\mathbf{1}_{Z>\zeta\vee\zeta_n}) &\ge c_2(\beta_n-\beta)^2,\\
(\beta_n-\alpha)^2\mathbb{Q}_n(\mathbf{1}_{\zeta_n<Z\le\zeta}) &\ge c_1\rho|\zeta-\zeta_n|, \quad \text{if } |\zeta-\zeta_n|\ge\frac{\delta^2}{8}>\frac{1}{\sqrt{m_n}},\\
(\alpha_n-\beta)^2\mathbb{Q}_n(\mathbf{1}_{\zeta<Z\le\zeta_n}) &\ge c_1\rho|\zeta-\zeta_n|, \quad \text{if } |\zeta-\zeta_n|\ge\frac{\delta^2}{8}>\frac{1}{\sqrt{m_n}},
\end{aligned}
$$

and therefore, noting that either $(\alpha-\alpha_n)^2+(\beta-\beta_n)^2\ge\frac{\delta^2}{8}$ or $|\zeta-\zeta_n|\ge\frac{\delta^2}{8}$, letting $c=\frac{1}{16}c_2\wedge(c_1\rho)$ and adding all the terms in the previous display, we get

$$
\sup_{\frac{\delta}{2}<d(\theta,\theta_n)<\delta}\{\mathcal{M}_n(\theta)-\mathcal{M}_n(\theta_n)\}\le\frac{8L^2}{\sqrt{m_n}}\delta-2c\delta^2 \quad \forall n\ge N.
$$

Hence, setting $\delta_n=\frac{8L^2}{c\sqrt{m_n}}\wedge\frac{2\sqrt{2}}{m_n^{1/4}}$ we get that

$$
\sup_{\frac{\delta}{2}<d(\theta,\theta_n)<\delta}\{\mathcal{M}_n(\theta)-\mathcal{M}_n(\theta_n)\}\le-c\delta^2 \quad \forall\ \delta_n\le\delta<\eta, \quad \forall n\ge N. \tag{5.29}
$$

Next we will show

$$
\sqrt{n}\mathbf{E}\left(\sup_{d(\theta,\theta_n)<\delta}\{|(\mathbb{M}_n-\mathcal{M}_n)(\theta)-(\mathbb{M}_n-\mathcal{M}_n)(\theta_n)|\}\right)\lesssim\frac{\sqrt{n}}{\sqrt{m_n}}\delta. \tag{5.30}
$$

Note that, using the expansion (5.27), $\mathbb{M}_n(\theta_n)=\mathcal{M}_n(\theta_n)=0$. To control the term $(\mathbb{M}_n-\mathcal{M}_n)(\theta)$ observe that it admits a very similar expansion as (5.28) with the $\mathbb{Q}_n$ replaced by $(\mathbb{P}_n^*-\mathbb{Q}_n)$; in particular, we can write the difference $\mathbb{M}_n(\theta)-\mathcal{M}_n(\theta)$ (by re-arranging the terms) as

$$
\begin{aligned}
&-2(\alpha_n-\alpha)(\mathbb{P}_n^*-\mathbb{Q}_n)(\tilde{\epsilon}_n\mathbf{1}_{Z\le\zeta\wedge\zeta_n})-2(\beta_n-\beta)(\mathbb{P}_n^*-\mathbb{Q}_n)(\tilde{\epsilon}_n\mathbf{1}_{Z>\zeta\vee\zeta_n})\\
&-2(\beta_n-\alpha)(\mathbb{P}_n^*-\mathbb{Q}_n)(\tilde{\epsilon}_n\mathbf{1}_{\zeta_n<Z\le\zeta})-2(\alpha_n-\beta)(\mathbb{P}_n^*-\mathbb{Q}_n)(\tilde{\epsilon}_n\mathbf{1}_{\zeta<Z\le\zeta_n})\\
&-(\alpha_n-\alpha)^2(\mathbb{P}_n^*-\mathbb{Q}_n)(\mathbf{1}_{Z\le\zeta\wedge\zeta_n})-(\beta_n-\beta)^2(\mathbb{P}_n^*-\mathbb{Q}_n)(\mathbf{1}_{Z>\zeta\vee\zeta_n})\\
&-(\alpha_n-\beta)^2(\mathbb{P}_n^*-\mathbb{Q}_n)(\mathbf{1}_{\zeta<Z\le\zeta_n})-(\beta_n-\alpha)^2(\mathbb{P}_n^*-\mathbb{Q}_n)(\mathbf{1}_{\zeta_n<Z\le\zeta})
\end{aligned} \tag{5.31}
$$

Each of these terms can be controlled by using Lemma 5.9.1 as

$$\begin{aligned}
\mathbf{E}\left(\left\|(\mathbb{P}_n^* - \mathbb{Q}_n)(\tilde{\epsilon}_n \mathbf{1}_{Z\le(\cdot)\wedge\zeta_n})\right\|_{[a,b]}\right) &\le \frac{J_{\mathscr{A}}(1)}{\sqrt{m_n}}\sqrt{\mathbb{Q}_n(\tilde{\epsilon}_n^2)} \\
\mathbf{E}\left(\left\|(\mathbb{P}_n^* - \mathbb{Q}_n)(\tilde{\epsilon}_n \mathbf{1}_{(\cdot)<Z\le\zeta_n})\right\|_{|\zeta-\zeta_n|<\delta^2}\right) &\le \frac{J_{\mathscr{A}}(1)}{\sqrt{m_n}}\sqrt{\mathbb{Q}_n(\tilde{\epsilon}_n^2 \mathbf{1}_{\zeta_n-\delta^2<Z\le\zeta_n+\delta^2})}.
\end{aligned}$$

Lemma 5.9.2 implies that $\mathbb{Q}_n(\tilde{\epsilon}_n^2 \mathbf{1}_{\zeta_n-\delta^2<Z\le\zeta_n+\delta^2}) \to \sigma^2\mathbb{P}(\zeta_0 - \delta^2 < Z \le \zeta_0 + \delta^2) = \sigma^2\{2f(\zeta_0)\delta^2 + o(\delta^2)\}$. Hence, there is a constant $R > 0$ such that the right side of the above equations are bounded by $R/\sqrt{m_n}$ and $R\sqrt{\delta^2 + o(\delta^2)}/\sqrt{m_n}$. Using similar arguments, we can in fact make R large enough so that the following inequalities hold too

$$\mathbf{E}\left(\left\|(\mathbb{P}_n^* - \mathbb{Q}_n)(\tilde{\epsilon}_n \mathbf{1}_{Z>(\cdot)\vee\zeta_n})\right\|_{[a,b]}\right) \le \frac{R}{\sqrt{m_n}} \tag{5.32}$$

$$\mathbf{E}\left(\left\|(\mathbb{P}_n^* - \mathbb{Q}_n)(\tilde{\epsilon}_n \mathbf{1}_{\zeta_n<Z\le(\cdot)})\right\|_{|\zeta-\zeta_n|<\delta^2}\right) \le \frac{R}{\sqrt{m_n}}\sqrt{\delta^2 + o(\delta^2)}. \tag{5.33}$$

We also assume that $R > J_{\mathcal{F}}(1)$. Using (5.32), (5.33), the discussion preceding the display, and grouping two consecutive terms at a time in the expansion (5.31), it is easily seen that

$$\begin{aligned}
&\sqrt{n}\mathbf{E}\left(\sup_{d(\theta,\theta_n)<\delta}\{|(\mathbb{M}_n - \mathcal{M}_n)(\theta) - (\mathbb{M}_n - \mathcal{M}_n)(\theta_n)|\}\right) \lesssim \frac{4R\sqrt{n}}{\sqrt{m_n}}\delta \\
&+\frac{4RL\sqrt{n}}{\sqrt{m_n}}\sqrt{\delta^2 + o(\delta^2)} + \frac{2R\sqrt{n}}{\sqrt{m_n}}\delta^2 + \frac{2RL^2 f(\zeta_0)\sqrt{n}}{\sqrt{m_n}}(\delta^2 + o(\delta^2)).
\end{aligned}$$

Thus by taking $\eta > 0$ small enough we can show that (5.30) holds for every $n \ge N$ and any $\delta \in [\delta_n, \eta)$, with δ_n and N defined as in (5.29). Defining $\phi_n(\delta) = \frac{\sqrt{n}}{\sqrt{m_n}}\delta$ and $r_n = \sqrt{m_n}$, the hypotheses of Theorem 3.4.1 of van der Vaart and Wellner (1996) are satisfied (note that Proposition 5.3.1 implies that $d(\theta_n, \theta_n^*) \xrightarrow{\mathbf{P}} 0$). Therefore, $r_n d(\theta_n, \theta_n^*) = \sqrt{m_n(\alpha_n^* - \alpha_n)^2 + m_n(\beta_n^* - \beta_n)^2} + \sqrt{m_n|\zeta_n^* - \zeta_n|} = O_{\mathbf{P}}(1)$. □

5.9.3 Proof of Lemma 5.3.1

Let $\eta > 0$ be an upper bound for the norm of the elements in K. The maximal inequality from Kim and Pollard (1990) and Lemma 5.9.1 imply

$$\begin{aligned}
\sqrt{m_n}\mathbf{E}\left(\left\|(\mathbb{P}_n^* - \mathbb{Q}_n)(\tilde{\epsilon}_n \mathbf{1}_{\zeta_n + \frac{(\cdot)}{m_n} < Z \le \zeta_n})\right\|_K\right) &\le J_{\mathscr{A}}(1)\sqrt{\mathbb{Q}_n(\tilde{\epsilon}_n^2 \mathbf{1}_{\zeta_n - \frac{\eta}{m_n} < Z \le \zeta_n})} \\
\sqrt{m_n}\mathbf{E}\left(\left\|(\mathbb{P}_n^* - \mathbb{Q}_n)(\mathbf{1}_{\zeta_n + \frac{(\cdot)}{m_n} < Z \le \zeta_n})\right\|_K\right) &\le J_{\mathcal{F}}(1)\sqrt{\mathbb{Q}_n(\mathbf{1}_{\zeta_n - \frac{\eta}{m_n} < Z \le \zeta_n})}.
\end{aligned}$$

By (i) and (iv) of Lemma 5.9.2 applied with $\mathbf{1}_{Z \le (\cdot) \wedge \zeta_n}$ in place of $\mathbf{1}_{(\cdot) < Z \le \zeta_n}$, we see that the righthand side of both the above inequalities go to zero. On the other hand, using (5.8) and (5.9) it is easy to see that both $\sqrt{m_n}\|\mathbb{Q}_n(\tilde{\epsilon}_n^2 \mathbf{1}_{\zeta_n + \frac{(\cdot)}{m_n} < Z \le \zeta_n})\|_K$ and $\sqrt{m_n}\|\mathbb{Q}_n(\mathbf{1}_{\zeta_n + \frac{(\cdot)}{m_n} < Z \le \zeta_n})\|_K$ converge to zero. Now, note that $\sqrt{m_n}\left\|\mathbb{P}_n^*\left(\tilde{\epsilon}_n \mathbf{1}_{\zeta_n + \frac{(\cdot)}{m_n} < Z \le \zeta_n}\right)\right\|_K$ is bounded by

$$\sqrt{m_n}\left\|(\mathbb{P}_n^* - \mathbb{Q}_n)(\tilde{\epsilon}_n \mathbf{1}_{\zeta_n + \frac{(\cdot)}{m_n} < Z \le \zeta_n})\right\|_K + \sqrt{m_n}\left\|\mathbb{Q}_n\left(|\tilde{\epsilon}_n| \mathbf{1}_{\zeta_n + \frac{(\cdot)}{m_n} < Z \le \zeta_n}\right)\right\|_K$$

and thus $\sqrt{m_n}\left\|\mathbb{P}_n^*\left(\tilde{\epsilon}_n \mathbf{1}_{\zeta_n + \frac{(\cdot)}{m_n} < Z \le \zeta_n}\right)\right\|_K \xrightarrow{\mathbb{L}_1} 0$. Similarly we can bound $\sqrt{m_n}\left\|\mathbb{P}_n^*\left(\mathbf{1}_{\zeta_n + \frac{(\cdot)}{m_n} < Z \le \zeta_n}\right)\right\|_K$ and show that it converges to zero in mean. Finally, from the expressions

$$\begin{aligned}
A_n^*(h_2) - \hat{A}_n(h_2, h_1) &= 2h_2\sqrt{m_n}\mathbb{P}_n^*\left(\tilde{\epsilon}_n \mathbf{1}_{\zeta_n + \frac{h_1}{m_n} < Z \le \zeta_n}\right) - h_2^2 \mathbb{P}_n^*\left(\mathbf{1}_{\zeta_n + \frac{(h_1}{m_n} < Z \le \zeta_n}\right), \\
C_n^*(h_1) - \hat{C}_n(h_3, h_1) &= 2h_3\sqrt{m_n}\mathbb{P}_n^*\left(\tilde{\epsilon}_n \mathbf{1}_{\zeta_n + \frac{h_1}{m_n} < Z \le \zeta_n}\right) \\
&\quad - \left(2h_3\sqrt{m_n}(\alpha_n - \beta_n) - h_3^2\right)\mathbb{P}_n^*\left(\mathbf{1}_{\zeta_n + \frac{h_1}{m_n} < Z \le \zeta_n}\right)
\end{aligned}$$

we get that $\left\|A_n^* - \hat{A}_n\right\|_K \xrightarrow{\mathbb{L}_1} 0$ and $\left\|C_n^* - \hat{C}_n\right\|_K \xrightarrow{\mathbb{L}_1} 0$. With completely analogous arguments, it is seen that $\left\|B_n^* - \hat{B}_n\right\|_K \xrightarrow{\mathbb{L}_1} 0$ and $\left\|D_n^* - \hat{D}_n\right\|_K \xrightarrow{\mathbb{L}_1} 0$ as well. Observing that $\hat{E}_n = \hat{A}_n + \hat{B}_n + \hat{C}_n + \hat{D}_n - \mathbb{P}_n^*(\tilde{\epsilon}_n^2)$ completes the proof of the result. $\square$

5.9.4 Proof of Lemma 5.3.2

It suffices to show that each of the components of $(\Xi_n)_{n=1}^{\infty}$ is tight. Write $\tilde{\epsilon}_{n,j} = \tilde{\epsilon}_n(Z_{n,j}, Y_{n,j})$ and let

$$\begin{aligned} r_n &= m_n \mathbb{Q}_n \left(e^{i\frac{\xi}{\sqrt{m_n}}\tilde{\epsilon}_n \mathbf{1}_{Z\le\zeta_n}} - 1 - i\frac{\xi}{m_n}\sqrt{m_n}\tilde{\epsilon}_n \mathbf{1}_{Z\le\zeta_n} + \frac{\xi^2}{2m_n}\tilde{\epsilon}_n^2 \mathbf{1}_{Z\le\zeta_n} \right) \\ &\le \frac{m_n^{-1/2}\xi^3 \mathbb{Q}_n|\tilde{\epsilon}_n|^3}{6}. \end{aligned}$$

Then, assumption (VIII) implies that $r_n \to 0$ as $n \to \infty$. Since the characteristic function of $\sqrt{m_n}\mathbb{P}_n^*(\tilde{\epsilon}_n \mathbf{1}_{Z\le\zeta_n})$ is given by

$$\mathbf{E}\left(e^{i\xi\sqrt{m_n}\mathbb{P}_n^*(\tilde{\epsilon}_n \mathbf{1}_{Z\le\zeta_n})}\right) = \left(1 + i\frac{\xi}{\sqrt{m_n}}\mathbb{Q}_n\left(\tilde{\epsilon}_n \mathbf{1}_{Z\le\zeta_n}\right) - \frac{\xi^2}{2m_n}\mathbb{Q}_n\left(\tilde{\epsilon}_n^2 \mathbf{1}_{Z\le\zeta_n}\right) + \frac{r_n}{m_n}\right)^{m_n}$$

taking the limit as $n \to \infty$ we can conclude that $\sqrt{m_n}\mathbb{P}_n^*(\tilde{\epsilon}_n \mathbf{1}_{Z\le\zeta_n}) \rightsquigarrow N(0, \mathbb{P}(Z \le \zeta_0)\sigma^2)$ by using (VII) and the fact that $(1 + \beta_n/n)^n \to e^{\beta}$ if $\beta_n \to \beta$. With similar arguments, it is seen that $\sqrt{m_n}\mathbb{P}_n^*(\tilde{\epsilon}_n \mathbf{1}_{Z>\zeta_n}) \rightsquigarrow N(0, \mathbb{P}(Z > \zeta_0)\sigma^2)$, so the first two components of the random vector of interest are uniformly tight.

Consider now the processes $\Gamma_n(t) = m_n\mathbb{P}_n^*(\mathbf{1}_{\zeta_n<Z\le\zeta_n+\frac{t}{m_n}})$ and $\Psi_n(t) = m_n\mathbb{P}_n^*(\tilde{\epsilon}_n\mathbf{1}_{\zeta_n<Z\le\zeta_n+\frac{t}{m_n}})$. For any process $\Psi \in \tilde{\mathcal{D}}_I$, $I \subset \mathbb{R}$ compact interval, $\delta > 0$, we write

$$w_{\Psi}^{''}(\delta) = \sup\left\{|\Psi(t_1) - \Psi(t)| \wedge |\Psi(t_2) - \Psi(t)|\right\}$$

where the supremum is taken over all $t_1 \le t \le t_2 \in I$ with $0 \le t_2 - t_1 \le \delta$. Also, for any $A \subset I$, define $w_{\Psi}(A) = \sup_{s,t\in A}\{|\Psi(t) - \Psi(s)|\}$. This agrees with the notation defined in Chapter 14 of Billingsley (1968). Let $\eta > 0$ be an upper bound for the absolute values of the elements of I, consider any $\rho > 0$, and define the numbers a_{Ψ}^{ρ} and a_{Γ}^{ρ} by,

$$\begin{aligned} a_{\Psi}^{\rho} &= \frac{1}{\rho}\sup_{n\in\mathbb{N}}\left\{m_n\mathbb{Q}_n\left(|\tilde{\epsilon}_n|\mathbf{1}_{\zeta_n<Z\le\zeta_n+\frac{\eta}{m_n}}\right)\right\} \\ a_{\Gamma}^{\rho} &= \frac{1}{\rho}\sup_{n\in\mathbb{N}}\left\{m_n\mathbb{Q}_n\left(\mathbf{1}_{\zeta_n<Z\le\zeta_n+\frac{\eta}{m_n}}\right)\right\}. \end{aligned}$$

Then, using Markov's inequality,

$$\varlimsup_{n\to\infty} \mathbf{P}\left(\sup_{t\in I}\{|\Psi_n(t)|\} > a_\Psi^\rho\right) \leq \rho \tag{5.34}$$

$$\varlimsup_{n\to\infty} \mathbf{P}\left(\sup_{t\in I}\{|\Gamma_n(t)|\} > a_\Gamma^\rho\right) \leq \rho. \tag{5.35}$$

Now, let $\rho, \gamma > 0$ be any pair of positive numbers and assume that $I = [a, b]$. Then, choose $\delta < \frac{\gamma}{8|b-a|f(\zeta_0)^2} \wedge \frac{|b-a|}{4} \wedge \frac{1}{f(\zeta_0)}$ so there is an integer $N \geq 2$ such that $\delta < \frac{|b-a|}{N} < 2\delta$. Define $s_j = a + \frac{j}{N}(b-a)$ and consider the partition $\{a = s_0 < s_1 < \ldots < s_N = b\}$ of I. Notice that if Ψ is a step function on I, for $w''_\Psi(\delta)$ to be positive, we need at least two jumps in an interval of size at most δ. Then, the probability that at least two jumps of the process Ψ_n happens on any interval $(s_{j-2}, s_j]$ is bounded from above by

$$\begin{aligned} a_{j,m_n} &:= \mathbf{P}\left(\bigcup_{1\leq k<l\leq m_n}\Big[m_n(Z_{n,k}-\zeta_n), m_n(Z_{n,l}-\zeta_n) \in (s_{j-2}, s_j]\Big]\right) \\ &\leq \frac{m_n^2}{2}\mathbb{Q}_n\left(\zeta_n + \frac{s_{j-2}}{m_n} < Z \leq \zeta_n + \frac{s_j}{m_n}\right)^2 \end{aligned}$$

and hence the limit superior of the probability that either Ψ_n or Γ_n has two jumps in any interval of the form $(s_{j-2}, s_j]$ is bounded from above by $2|b-a|^2 f(\zeta_0)^2/N^2$ by (VI). Therefore, the probability that at least two jumps happen in any interval of size at most δ is asymptotically bounded from above by

$$\sum_{i=2}^{N} a_{j,m_n} \leq \sum_{i=2}^{N} 2|b-a|^2 f(\zeta_0)^2/N^2 \leq 4(N-1)f(\zeta_0)^2|b-a|\delta/N \leq \gamma.$$

Thus,

$$\varlimsup_{n\to\infty} \mathbf{P}\left(w''_{\Psi_n}(\delta) > \rho\right) < \gamma \tag{5.36}$$

The exact same argument can be used to show that

$$\varlimsup_{n\to\infty} \mathbf{P}\left(w''_{\Gamma_n}(\delta) > \rho\right) < \gamma. \tag{5.37}$$

Now, note that

$$\begin{aligned}
\mathbf{P}\left(w_{\Psi_n}\left([a,a+\delta)\right)>\rho\right) &\leq \mathbf{P}\left(\bigcup_{j=1}^{m_n} m_n(Z_{n,j}-\zeta_n)\in[a,a+\delta)>\rho\right)\\
&\leq m_n\mathbb{Q}_n\left(\zeta_n+\frac{a}{m_n}<Z\leq\zeta_n+\frac{a+\delta}{m_n}\right)
\end{aligned}$$

which implies that

$$\overline{\lim_{n\to\infty}}\,\mathbf{P}\left(w_{\Psi_n}\left([a,a+\delta)\right)>\rho\right)\leq\delta f(\zeta_0)<\gamma. \tag{5.38}$$

A similar analysis leads to the following bounds

$$\overline{\lim_{n\to\infty}}\,\mathbf{P}\left(w_{\Psi_n}\left([b-\delta,b)\right)>\rho\right) < \gamma \tag{5.39}$$

$$\overline{\lim_{n\to\infty}}\,\mathbf{P}\left(w_{\Gamma_n}\left([a,a+\delta)\right)>\rho\right) < \gamma \tag{5.40}$$

$$\overline{\lim_{n\to\infty}}\,\mathbf{P}\left(w_{\Gamma_n}\left([b-\delta,b)\right)>\rho\right) < \gamma. \tag{5.41}$$

Putting together (5.34), (5.35), (5.36), (5.37), (5.38), (5.39), (5.40) and (5.41) and using Theorem 15.3 of Billingsley (1968) we obtain that both sequences $(\Psi_n)_{n=1}^{\infty}$ and $(\Gamma_n)_{n=1}^{\infty}$ are uniformly tight in $\tilde{\mathcal{D}}_I$. Similar arguments show the tightness of the third and fourth components of the process. Therefore, $(\Xi_n)_{n=1}^{\infty}$ is uniformly tight. The uniform tightness of $(E_n^*)_{n=1}^{\infty}$ now follows from the fact that $(\Xi_n)_{n=1}^{\infty}$ is uniformly tight and E_n^* is a continuous function of Ξ_n. $\square$

5.9.5 Proof of Lemma 5.3.3

In view of Lemma 5.3.2, to show (i) it suffices to show convergence of the finite dimensional distributions. To this end, consider the real numbers $t_{-N_-} <$

$\ldots < t_{-1} < 0 = t_0 < t_1 < \ldots < t_{N_+}$ and the linear combination

$$\begin{aligned} W_n &= \mu\sqrt{m_n}\mathbb{P}_n^*(\tilde{\epsilon}_n\mathbf{1}_{Z\leq\zeta_n}) + \lambda\sqrt{m_n}\mathbb{P}_n^*\left(\tilde{\epsilon}_n\mathbf{1}_{Z>\zeta_n}\right) \\ &\quad + \sum_{j=1}^{N_-}\left\{\xi_{-j}m_n\mathbb{P}_n^*\left(\tilde{\epsilon}_n\mathbf{1}_{\zeta_n+\frac{t_{-j}}{m_n}<Z\leq\zeta_n}\right) + \eta_{-j}m_n\mathbb{P}_n^*\left(\mathbf{1}_{\zeta_n+\frac{t_{-j}}{m_n}<Z\leq\zeta_n}\right)\right\} \\ &\quad + \sum_{j=1}^{N_+}\left\{\xi_j m_n\mathbb{P}_n^*\left(\tilde{\epsilon}_n\mathbf{1}_{\zeta_n<Z\leq\zeta_n+\frac{t_j}{m_n}}\right) + \eta_j m_n\mathbb{P}_n^*\left(\mathbf{1}_{\zeta_n<Z\leq\zeta_n+\frac{t_j}{m_n}}\right)\right\} \end{aligned} \tag{5.42}$$

where μ, λ and the ξ_j's and the η_j's are arbitrary real numbers. Now, set $\xi_0 = \eta_0 = 0$ and define

$$\mu_{\pm j} = \sum_{k=j}^{N_\pm}\eta_{\pm k} \quad \text{and} \quad \lambda_{\pm j} = \sum_{k=j}^{N_\pm}\xi_{\pm k}. \tag{5.43}$$

Then grouping terms appropriately we can rewrite W_n as

$$\begin{aligned} W_n &= \mu\sqrt{m_n}\mathbb{P}_n^*\left(\tilde{\epsilon}_n\mathbf{1}_{Z\leq\zeta_n+\frac{t_{-N_-}}{m_n}}\right) + \lambda\sqrt{m_n}\mathbb{P}_n^*\left(\tilde{\epsilon}_n\mathbf{1}_{Z>\zeta_n+\frac{t_{N_+}}{m_n}}\right) \\ &\quad + \sum_{j=1}^{N_-}(\lambda_{-j}m_n + \mu\sqrt{m_n})\mathbb{P}_n^*\left(\tilde{\epsilon}_n\mathbf{1}_{\zeta_n+\frac{t_{-j}}{m_n}<Z\leq\zeta_n+\frac{t_{-j+1}}{m_n}}\right) \\ &\quad + \sum_{j=1}^{N_-}\mu_{-j}m_n\mathbb{P}_n^*\left(\mathbf{1}_{\zeta_n+\frac{t_{-j}}{m_n}<Z\leq\zeta_n+\frac{t_{-j+1}}{m_n}}\right) \\ &\quad + \sum_{j=1}^{N_+}(\lambda_j m_n + \lambda\sqrt{m_n})\mathbb{P}_n^*\left(\tilde{\epsilon}_n\mathbf{1}_{\zeta_n+\frac{t_{j-1}}{m_n}<Z\leq\zeta_n+\frac{t_j}{m_n}}\right) \\ &\quad + \sum_{j=1}^{N_+}\mu_j m_n\mathbb{P}_n^*\left(\mathbf{1}_{\zeta_n+\frac{t_{j-1}}{m_n}<Z\leq\zeta_n+\frac{t_j}{m_n}}\right). \end{aligned}$$

Using the independence of $X_{n,1},\ldots,X_{n,m_n}$, the characteristic function of W_n is

$$\begin{aligned} \mathbf{E}\left(e^{isW_n}\right) = \Bigg[1 &+ \sum_{j=1}^{N_-}\mathbb{Q}_n\left((e^{is(\frac{\mu}{\sqrt{m_n}}+\lambda_{-j})\tilde{\epsilon}_n+is\mu_{-j}} - 1)\mathbf{1}_{\zeta_n+\frac{t_{-j}}{m_n}<Z\leq\zeta_n+\frac{t_{-j+1}}{m_n}}\right) \\ &+ \mathbb{Q}_n\left((e^{i\frac{s\mu}{\sqrt{m_n}}\tilde{\epsilon}_n} - 1)\mathbf{1}_{Z\leq\zeta_n+\frac{t_{-N_-}}{m_n}}\right) + \mathbb{Q}_n\left((e^{i\frac{s\lambda}{\sqrt{m_n}}\tilde{\epsilon}_n} - 1)\mathbf{1}_{Z>\zeta_n+\frac{t_{N_+}}{m_n}}\right) \\ &+ \sum_{j=1}^{N_+}\mathbb{Q}_n\left((e^{is(\frac{\lambda}{\sqrt{m_n}}+\lambda_j)\tilde{\epsilon}_n+is\mu_j} - 1)\mathbf{1}_{\zeta_n+\frac{t_{j-1}}{m_n}<Z\leq\zeta_n+\frac{t_j}{m_n}}\right)\Bigg]^{m_n} \end{aligned} \tag{5.44}$$

Let r_n be given by

$$r_n = m_n \mathbb{Q}_n \left[\left(e^{i\frac{s\mu}{\sqrt{m_n}}\tilde{\epsilon}_n} - 1 - i\frac{s\mu}{\sqrt{m_n}}\tilde{\epsilon}_n + \frac{s^2\mu^2}{2m_n}\tilde{\epsilon}_n^2 \right) \mathbf{1}_{Z \leq \zeta_n + \frac{t_{-N_-}}{m_n}} \right] \leq \frac{s^3 \mathbb{Q}_n |\tilde{\epsilon}_n^3|}{6\sqrt{m_n}}.$$

Condition (VIII) now implies that $r_n = o(1)$. But note that

$$\mathbb{Q}_n \left((e^{i\frac{s\mu}{\sqrt{m_n}}\tilde{\epsilon}_n} - 1)\mathbf{1}_{Z \leq \zeta_n + \frac{t_{-N_-}}{m_n}} \right) = i\frac{s\mu}{m_n}\sqrt{m_n}\mathbb{Q}_n \left(\tilde{\epsilon}\mathbf{1}_{Z \leq \zeta_n + \frac{t_{-N_-}}{m_n}} \right)$$
$$- \frac{s^2\mu^2}{2m_n}\mathbb{Q}_n \left(\tilde{\epsilon}_n^2 \mathbf{1}_{Z \leq \zeta_n + \frac{t_{-N_-}}{m_n}} \right) + \frac{r_n}{m_n}$$

and so (i) of Lemma 5.9.2 together with condition (VII) and (5.8) imply that

$$m_n \mathbb{Q}_n \left((e^{i\frac{s\mu}{\sqrt{m_n}}\tilde{\epsilon}_n} - 1)\mathbf{1}_{Z \leq \zeta_n + \frac{t_{-N_-}}{m_n}} \right) = -\frac{s^2\mu^2}{2}\sigma^2 \mathbb{P}(Z \leq \zeta_0) + o(1). \quad (5.45)$$

Following a completely analogous argument one can show that

$$m_n \mathbb{Q}_n \left(\left(e^{i\frac{s\lambda}{\sqrt{m_n}}\tilde{\epsilon}_n} - 1 \right) \mathbf{1}_{Z > \zeta_n + \frac{t_{N_+}}{m_n}} \right) = -\frac{s^2\lambda^2}{2}\sigma^2 \mathbb{P}(Z > \zeta_0) + o(1). \quad (5.46)$$

Now, take $1 \leq j \leq N_+$, and observe that equation (5.8) implies

$$m_n \left| \mathbb{Q}_n \left((e^{is(\frac{\lambda}{\sqrt{m_n}} + \lambda_j)\tilde{\epsilon}_n + is\mu_j} - e^{is\lambda_j\tilde{\epsilon}_n + is\mu_j})\mathbf{1}_{\zeta_n + \frac{t_{j-1}}{m_n} < Z \leq \zeta_n + \frac{t_j}{m_n}} \right) \right|$$
$$\leq |s\lambda|\sqrt{m_n}\mathbb{Q}_n \left(|\tilde{\epsilon}_n| \mathbf{1}_{\zeta_n + \frac{t_{j-1}}{m_n} < Z \leq \zeta_n + \frac{t_j}{m_n}} \right) \to 0.$$

Using (VI) we can write

$$m_n \mathbb{Q}_n \left((e^{is(\frac{\lambda}{\sqrt{m_n}} + \lambda_j)\tilde{\epsilon}_n + is\mu_j} - 1)\mathbf{1}_{\zeta_n + \frac{t_{j-1}}{m_n} < Z \leq \zeta_n + \frac{t_j}{m_n}} \right)$$
$$= (\varphi(s\lambda_j)e^{is\mu_j} - 1) f(\zeta_0)(t_j - t_{j-1}) + o(1)$$

where φ is the characteristic function of ϵ (under $\mathbb{P}$). Thus,

$$m_n \sum_{j=1}^{N_+} \mathbb{Q}_n \left((e^{is(\frac{\lambda}{\sqrt{m_n}} + \lambda_j)\tilde{\epsilon}_n + is\mu_j} - 1)\mathbf{1}_{\zeta_n + \frac{t_{j-1}}{m_n} < Z \leq \zeta_n + \frac{t_j}{m_n}} \right)$$
$$= \sum_{j=1}^{N_+} (t_j - t_{j-1}) f(\zeta_0)(\varphi(s\lambda_j)e^{is\mu_j} - 1) + o(1). \quad (5.47)$$

Similarly, one can prove that

$$m_n \sum_{j=1}^{N_-} \mathbb{Q}_n \left((e^{is(\frac{\mu}{\sqrt{m_n}}+\lambda_{-j})\tilde{\epsilon}_n + is\mu_{-j}} - 1)\mathbf{1}_{\zeta_n + \frac{t_{-j}}{m_n} < Z \le \zeta_n + \frac{t_{-j+1}}{m_n}} \right)$$

$$= \sum_{j=1}^{N_-} (t_{-j+1} - t_{-j}) f(\zeta_0)(\varphi(s\lambda_{-j})e^{is\mu_{-j}} - 1) + o(1). \tag{5.48}$$

So putting (5.42), (5.43), (5.44), (5.45), (5.46), (5.47) and (5.48) together we see that,

$$\begin{aligned} \mathbf{E}\left(e^{isW_n}\right) \quad \to \quad & \exp\left[\sum_{j=1}^{N_-} f(\zeta_0)(t_{-j+1} - t_{-j}) \left\{ \varphi\left(s(\sum_{k=j}^{N_-} \xi_{-k}) \right) e^{is\sum_{k=j}^{N_-} \eta_{-k}} - 1 \right\} \right. \\ & - \frac{s^2\mu^2\sigma^2}{2}\mathbb{P}(Z \le \zeta_0) - \frac{s^2\lambda^2\sigma^2}{2}\mathbb{P}(Z > \zeta_0) \\ & \left. + \sum_{j=1}^{N_+} f(\zeta_0)(t_j - t_{j-1}) \left\{ \varphi\left(s(\sum_{k=j}^{N_+} \xi_k) \right) e^{is\left(\sum_{k=j}^{N_+} \eta_k\right)} - 1 \right\} \right] \end{aligned} \tag{5.49}$$

But the right-hand side of (5.49) is precisely $\mathbf{E}\left(e^{isW}\right)$ where, with the notation of (5.11), W is given by

$$\begin{aligned} W = \mu\mathbf{Z}_1 + \lambda\mathbf{Z}_2 + \sum_{k=1}^{N_-} & \left(\xi_{-k} \sum_{0<j\le\nu_1(-t_{-k})} v_k \mathbf{1}_{t_{-k}<0} + \eta_{-k}\nu_1(-t_{-k})\mathbf{1}_{t_{-k}<0} \right) \\ & + \sum_{k=1}^{N_+} \left(\xi_k \sum_{0<j\le\nu_2(t_k)} u_k \mathbf{1}_{t_k\ge 0} + \eta_k\nu_2(t_k)\mathbf{1}_{t_k\ge 0} \right) \end{aligned}$$

and thus $W_n \rightsquigarrow W$. From the fact that μ, λ, the ξ_j's and the η_j's were arbitrarily chosen, by the Cramer-Wold device

$$\left(\Xi_n(t_{-N_-}), \ldots, \Xi_n(t_{-1}), \Xi_n(t_1), \ldots, \Xi_n(t_{N_+})\right)' \rightsquigarrow \left(\Xi(t_{-N_-}), \ldots, \Xi(t_{-1}), \Xi(t_1), \ldots, \Xi(t_{N_+})\right)'.$$

This gives the convergence of the finite dimensional distributions, proving (i). An application of the continuous mapping theorem shows that (i) implies (ii). Further, Lemma 5.3.1 and (ii) now imply (iii). $\square$

5.9.6 Proof of Lemma 5.3.4

Every sample path of $E^* = E^*(h_1, h_2, h_3)$ can be written as

$$2h_2\mathbf{Z}_1 - h_2^2\mathbb{P}(Z \le \zeta_0) + 2h_3\mathbf{Z}_2 - h_3^2\mathbb{P}(Z > \zeta_0) + \mathbf{1}_{h_1<0}2(\alpha_0 - \beta_0)\sum_{j=1}^{\nu_1(-h_1)} v_j$$

$$-(\alpha_0 - \beta_0)^2\nu_1(-h_1)\mathbf{1}_{h_1<0} + \mathbf{1}_{h_1\ge 0}2(\beta_0 - \alpha_0)\sum_{j=1}^{\nu_2(h_1)} u_j - \mathbf{1}_{h_1\ge 0}(\alpha_0 - \beta_0)^2\nu_2(h_1).$$

From this last expression it is obvious that for any fixed h_1, the $E^*(h_1, \cdot, \cdot)$ gets maximized at $\phi_2^* = \mathbf{Z}_1/\mathbb{P}(Z \le \zeta_0)$ and $\phi_3^* = \mathbf{Z}_2/\mathbb{P}(Z > \zeta_0)$. The independence of the three co-ordinates follows from the fact that ϕ_2^* depends only on $\mathbf{Z}_1$, ϕ_3^* depends only on $\mathbf{Z}_2$, and ϕ_1^* depends only on $\mathbf{u}$, $\mathbf{v}$, ν_1 and ν_2. Since E^* is piecewise constant in the third argument h_3, to complete the proof it is enough to show that $E^*(h_1, \phi_1^*, \phi_2^*) \to -\infty$ as $|h_1| \to \infty$. But this follows from the law of the iterated logarithm (applied to the random walks defined by the v_i's and u_i's) together with the fact that $\nu_1(t) \wedge \nu_2(t) \xrightarrow{a.s.} \infty$ as $t \to \infty$. Note that $\sum_{j=1}^{\nu_1(-h_1)} v_j$ and $\sum_{j=1}^{\nu_2(h_1)} u_j$ are of orders $O(\sqrt{\nu_1 \log\log \nu_1})$ and $O(\sqrt{\nu_2 \log\log \nu_2})$ a.s. as $h_1 \to -\infty$ and $h_1 \to \infty$, respectively. $\square$

5.9.7 Proof of Proposition 5.3.3

Lemma 5.3.4 and the fact that the u_i's and the v_i's come from a continuous distribution, show that (E^*, J^*) satisfy the hypotheses of Theorem 4.3.2, and in particular that (4.5) holds. Moreover, Proposition 5.3.2 shows that the sequence $(m_n(\zeta_n^* - \zeta_n), \sqrt{m_n}(\alpha_n^* - \alpha_n), \sqrt{m_n}(\beta_n^* - \beta_n))'$ is tight. The result now follows from a direct application of Theorem 4.3.2. $\square$

5.9.8 Proof of Lemma 5.4.1

We expand $m_\theta(X)$ as in (5.26) but with $\epsilon = Y - \alpha_0 \mathbf{1}_{Z\le\zeta_0} - \beta_0 \mathbf{1}_{Z>\zeta_0}$ in place of $\tilde{\epsilon}_n$ to get

$$\begin{aligned} m_\theta(X) &= -(\epsilon+\alpha_0-\alpha)^2\mathbf{1}_{Z\le\zeta_0\wedge\zeta} - (\epsilon+\beta_0-\alpha)^2\mathbf{1}_{\zeta_0<Z\le\zeta} \\ &\quad -(\epsilon+\alpha_0-\beta)^2\mathbf{1}_{\zeta<Z\le\zeta_0} - (\epsilon+\beta_0-\beta)^2\mathbf{1}_{Z>\zeta_0\vee\zeta}. \end{aligned} \tag{5.50}$$

Letting $\hat{\gamma}_n = (\hat{\alpha}_n, \hat{\beta}_n)$, we can also bound $M_n(\theta_0)$ using a similar argument as in the proof of Proposition 5.3.1 to obtain

$$\begin{aligned} &|\hat{\gamma}_n - \gamma_0|^2 \mathbb{P}_n(Z<a)\wedge\mathbb{P}_n(Z>b) \\ \le\ & \mathbb{P}_n\left(\epsilon^2\mathbf{1}_{a\le Z\le b}\right) + 2|\hat{\gamma}_n-\gamma_0|\left(|\mathbb{P}_n\left(\epsilon\mathbf{1}_{Z<a}\right)| + |\mathbb{P}_n\left(\epsilon\mathbf{1}_{Z>b}\right)|\right). \end{aligned}$$

By the strong law of large numbers

$$\begin{aligned} \mathbb{P}_n(Z<a)\wedge\mathbb{P}_n(Z>b) &\xrightarrow{a.s.} \mathbb{P}(Z<a)\wedge\mathbb{P}(Z>b) \\ \mathbb{P}_n\left(\epsilon^2\mathbf{1}_{a\le Z\le b}\right) &\xrightarrow{a.s.} \sigma^2\mathbb{P}\left(a\le Z\le b\right) \quad \text{and} \\ |\mathbb{P}_n\left(\epsilon\mathbf{1}_{Z<a}\right)| + |\mathbb{P}_n\left(\epsilon\mathbf{1}_{Z>b}\right)| &\xrightarrow{a.s.} 0. \end{aligned}$$

Therefore, w.p. 1 we can write

$$|\hat{\gamma}_n - \gamma_0|^2 \le O(1) + |\hat{\gamma}_n-\gamma_0| o(1)$$

and thus the sequence $(\hat{\gamma}_n - \gamma_0)_{n=1}^\infty$ is bounded w.p. 1.

Now, take any compact set $K \subset \Theta$ and consider the classes of functions

$$\begin{aligned} \mathcal{K}_1 &= \left\{(\epsilon+\alpha_0-\alpha)^2\,\mathbf{1}_{(-\infty,\zeta\wedge\zeta_0]}\right\}_{\theta\in K} \\ \mathcal{K}_2 &= \left\{(\epsilon+\beta_0-\alpha)^2\,\mathbf{1}_{(\zeta_0,\zeta]}\right\}_{\theta\in K} \\ \mathcal{K}_3 &= \left\{(\epsilon+\alpha_0-\beta)^2\,\mathbf{1}_{(\zeta,\zeta_0]}\right\}_{\theta\in K} \\ \mathcal{K}_4 &= \left\{(\epsilon+\beta_0-\beta)^2\,\mathbf{1}_{(\zeta\vee\zeta_0,\infty)}\right\}_{\theta\in K}. \end{aligned}$$

If A^* is an upper bound for the norm of the elements in K, we can see that each of these classes is a VC-subgraph class with integrable envelope

$(|\epsilon| + A^* + |\gamma_0|)^2$. With the notation $\|Q\|_{\mathcal{F}} = \sup\{|Qf| : f \in \mathcal{F}\}$ for classes of functions $\mathcal{F}$ and probability measures Q, a combination of Theorems 2.6.7 and 2.4.3 of van der Vaart and Wellner (1996) shows that all four quantities $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{K}_j}$, $j = 1, 2, 3, 4$, converge to zero almost surely. Therefore using (5.50), we get the inequality

$$\|M_n - M\|_K \leq \sum_{1 \leq j \leq 4} \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{K}_j}$$

which now implies (i) (Since $M_n, M \in \mathcal{D}_K$, $\|M_n - M\|_K$ is measurable.). The second assertion follows immediately from (ii).

Consider a family of compact rectangles $\Theta_n \subset \Theta_{n+1}$ such that $\Theta = \cup_{n=1}^{\infty} \Theta_n$. Then, since the sequence $(\hat{\gamma}_n - \gamma_0)_{n=1}^{\infty}$ is almost surely bounded, w.p. 1 we have that there is some $m \in \mathbb{N}$ such that Θ_m contains both θ_0 and the entire sequence $(\hat{\theta}_n)_{n=1}^{\infty}$. Finally, from (5.27) with θ_n replaced by θ_0 it is seen that

$$\begin{aligned} M(\theta) = -\sigma^2 - (\alpha_0 - \alpha)^2 \mathbb{P}(Z \leq \zeta \wedge \zeta_0) - (\alpha_0 - \beta)^2 \mathbb{P}(\zeta < Z \leq \zeta_0) \\ -(\alpha - \beta_0)^2 \mathbb{P}(\zeta_0 < Z \leq \zeta) - (\beta_0 - \beta)^2 \mathbb{P}(Z > \zeta \vee \zeta_0). \end{aligned}$$

As $\alpha_0 \neq \beta_0$ and Z has a strictly positive density on $[a, b]$, the last equation shows that M satisfies the conditions of Lemma 4.2.9. Since the event that $M_n \to M$ in $\mathcal{D}_{\Theta_k}$ for all $k \in \mathbb{N}$ has probability one, Lemma 4.2.9 allows us to conclude that $\text{sargmax}(M_n) = \hat{\theta}_n \xrightarrow{a.s.} \theta_0$. $\square$

5.9.9 Proof of Lemma 5.4.2

Let $\rho, \delta > 0$. We know from Corollary 5.3.1 that the sequences $(\sqrt{n}(\hat{\alpha}_n - \alpha_0))_{n=1}^{\infty}$, $\left(n(\hat{\zeta}_n - \zeta_0)\right)_{n=1}^{\infty}$ and $\left(n\mathbb{P}_n\left(\zeta_0 - \frac{h}{n} < Z \leq \zeta_0 + \frac{h}{n}\right)\right)_{n=1}^{\infty}$, for any $h > 0$, are all stochastically bounded. Thus, since $m_n = O(n)$ there is $L > 0$ such that

$\mathbf{P}\left(m_n|\hat{\zeta}_n-\zeta_0|>L\right)<\rho$ and $\mathbf{P}\left(\sqrt{m_n}|\hat{\alpha}_n-\alpha_0|>L\right)<\rho$ for any $n\in\mathbb{N}$. Therefore,

$$\begin{aligned}
&\mathbf{P}\left(m_n^\gamma\left\|\mathbb{P}_n(\hat{\zeta}_n+\frac{(\cdot)}{m_n}<Z\leq\hat{\zeta}_n)\right\|_K>\delta\right)\\
\leq\ &\frac{m_n^\gamma}{\delta}\mathbf{E}\left(\mathbb{P}_n\left(\zeta_0-\frac{L+\eta}{m_n}<Z\leq\zeta_0+\frac{L}{m_n}\right)\right)+\mathbf{P}\left(m_n|\hat{\zeta}_n-\zeta_0|>L\right)\\
\leq\ &f(\zeta_0)\frac{\eta+2L}{\delta}m_n^{\gamma-1}+o\left(m_n^{\gamma-1}\right)+\rho,
\end{aligned}$$

so by letting $n\to\infty$ and then $\rho\to 0$ we get (i).

We prove (ii) for when $p=1$, the case $p=2$ follows from similar arguments. Note that if $m_n|\hat{\zeta}_n-\zeta_0|\leq L$, then $m_n^\gamma\|\mathbb{P}_n(|\tilde{\epsilon}_n|\mathbf{1}_{\hat{\zeta}_n+\frac{(\cdot)}{m_n}<Z\leq\hat{\zeta}_n})\|_K$ can be bounded by

$$m_n^\gamma\left\|\mathbb{P}_n\left(|\epsilon|\mathbf{1}_{\zeta_0-\frac{L}{m_n}+\frac{(\cdot)}{m_n}<Z\leq\zeta_0+\frac{L}{m_n}}\right)\right\|_K+m_n^\gamma|\hat{\alpha}_n-\alpha_0|\left\|\mathbb{P}_n(\zeta_0-\frac{L}{m_n}+\frac{(\cdot)}{m_n}<Z\leq\zeta_0+\frac{L}{m_n})\right\|_K.$$

But just as in the proof of (i), we have

$$\begin{aligned}
&\mathbf{P}\left(m_n^\gamma\left\|\mathbb{P}_n(|\epsilon|\mathbf{1}_{\hat{\zeta}_n+\frac{(\cdot)}{m_n}<Z\leq\hat{\zeta}_n})\right\|_K>\delta\right)\\
\leq\ &\mathbf{P}\left(m_n^\gamma\left\|\mathbb{P}_n\left(|\epsilon|\mathbf{1}_{\zeta_0-\frac{L}{m_n}+\frac{(\cdot)}{m_n}<Z\leq\zeta_0+\frac{L}{m_n}}\right)\right\|_K>\frac{\delta}{2}\right)+\\
&\mathbf{P}\left(m_n^\gamma|\hat{\alpha}_n-\alpha_0|\mathbb{P}_n(\zeta_0-\frac{L}{m_n}+\frac{\eta}{m_n}<Z\leq\zeta_0+\frac{L}{m_n})>\frac{\delta}{2}\right)+\mathbf{P}\left(m_n|\hat{\zeta}_n-\zeta_0|>L\right)\\
\leq\ &\frac{2m_n^\gamma}{\delta}\mathbf{E}\left(\mathbb{P}_n\left(|\epsilon|\mathbf{1}_{\zeta_0-\frac{L}{m_n}+\frac{\eta}{m_n}<Z\leq\zeta_0+\frac{L}{m_n}}\right)\right)+\\
&\mathbf{P}\left(m_n^\gamma|\hat{\alpha}_n-\alpha_0|\mathbb{P}_n(\zeta_0-\frac{L}{m_n}+\frac{\eta}{m_n}<Z\leq\zeta_0+\frac{L}{m_n})>\frac{\delta}{2}\right)+\mathbf{P}\left(m_n|\hat{\zeta}_n-\zeta_0|>L\right)\\
\leq\ &f(\zeta_0)\mathbf{E}\left(|\epsilon|\right)\frac{2(\eta+2L)}{\delta}m_n^{\gamma-1}+o\left(m_n^{\gamma-1}\right)+\\
&\mathbf{P}\left(m_n^\gamma|\hat{\alpha}_n-\alpha_0|\mathbb{P}_n(\zeta_0-\frac{L}{m_n}+\frac{\eta}{m_n}<Z\leq\zeta_0+\frac{L}{m_n})>\frac{\delta}{2}\right)+\rho.
\end{aligned}$$

The result follows again by letting $n\to\infty$ and $\rho\to 0$. $\square$

The next results will be useful to support our conjecture of inconsistency of some of our bootstrap scenarios.

Lemma 5.9.4 *Let $\lambda, B>0$, $\rho\in(0,\frac{1}{2})$ and H_λ be the distribution function of a Poisson random variable with mean λ. For each value of λ write $L_{\lambda+B}^\rho=$*

$\min\{n \in \mathbb{N} : H_{\lambda+B}(n) > \rho\}$ *and* $U_\lambda^\rho = \max\{n \in \mathbb{N} : 1 - H_\lambda(n) > \rho\}$. *Then, there is* $\lambda_* > 0$ *such that* $L_{\lambda+B}^\rho < U_\lambda^\rho$ *for all* $\lambda \geq \lambda_*$.

Proof: Let c_λ be the median (i.e. $c_\lambda = \min\{n \in \mathbb{N} : H_\lambda(n) > \frac{1}{2}\}$.) of H_λ. Observe that $c_\lambda \leq U_\lambda^\rho$. According to Hazma (1995), $|c_\lambda - \lambda| < \log(2)$ for any positive λ. Letting $\lfloor x \rfloor$ denote the greatest integer less than or equal to x, we have

$$\begin{aligned} & |H_{\lambda+B}(c_{\lambda+B}) - H_{\lambda+B}(c_\lambda)| \\ \leq \; & |H_{\lambda+B}(\lambda + B + \log(2)) - H_{\lambda+B}(\lambda - \log(2))| \\ \leq \; & (B + 2\log(2))e^{-(\lambda+B)}\frac{(\lambda+B)^{\lfloor \lambda+B \rfloor}}{\lfloor \lambda + B \rfloor !} \to 0 \text{ as } \lambda \to \infty. \end{aligned}$$

as the Poisson mass function has a maximum at $\lfloor \lambda + B \rfloor$. Therefore, $\underline{\lim}_{\lambda\to\infty} H_{\lambda+B}(U_\lambda^\rho) \geq 1/2$. But we also note that $\sup_{n\in\mathbb{N}}\{H_{\lambda+B}(n+1) - H_{\lambda+B}(n)\} \to 0$ as $\lambda \to \infty$. Thus,

$$\overline{\lim_{\lambda\to\infty}} H_{\lambda+B}(L_{\lambda+B}^\rho + 1) = \rho < \frac{1}{2} \leq \underline{\lim_{\lambda\to\infty}} H_{\lambda+B}(U_\lambda^\rho).$$

It follows that $U_\lambda^\rho > L_{\lambda+B}^\rho$ for all λ sufficiently large. $\square$

Lemma 5.9.5 *Let* $\lambda, B > 0$, $0 < \rho < \frac{1}{2}$, μ *and* ν *be two nondegenerate Borel probability measures on* $\mathbb{R}$ *and* $H_{\mu,\lambda}$ *denote the compound Poisson distribution with intensity* λ *and compounding distribution* μ. *For each value of* λ *write* $L_{\nu,\lambda+B}^\rho = \inf\{s \in \mathbb{R} : H_{\nu,\lambda+B}(s) \geq \rho\}$ *and* $U_{\mu,\lambda}^\rho = \sup\{s \in \mathbb{R} : 1 - H_{\mu,\lambda}(s) \geq \rho\}$. *In addition, assume that* $\int x^2\nu(dx), \int x^2\mu(dx) < \infty$ *and that* $\int x\nu(dx) \leq \int x\mu(dx)$. *Then there is* $\lambda_* > 0$ *such that* $L_{\nu,\lambda+B}^\rho < U_{\mu,\lambda}^\rho$ *for all* $\lambda \geq \lambda_*$. *Moreover, let* $0 < r < 1$, *suppose that there is another Borel probability measure* γ *on* $\mathbb{R}$, *also satisfying* $\int x^2\gamma(dx) < \infty$, *and define* $\nu_\gamma := \frac{rB}{\lambda+B}\gamma + \frac{\lambda+(1-r)B}{\lambda+B}\mu$ *with its corresponding constant* $L_{\nu_\gamma,\lambda+B}^\rho = \inf\left\{s \in \mathbb{R} : H_{\nu_\gamma,\lambda+B}(s) \geq \rho\right\}$. *Then there is* $\lambda_* > 0$ *such that* $L_{\nu_\gamma,\lambda+B}^\rho < U_{\mu,\lambda}^\rho$ *for all* $\lambda \geq \lambda_*$.

Proof: Denote by Φ the standard normal distribution and $\mathbf{z}_\alpha$ the lower α-quantile of Φ (i.e. $\Phi(\mathbf{z}_\alpha) = \alpha$). Also, write $c_\mu := \int x\mu(dx)$, $d_\mu := \int x^2\mu(dx)$ and define the corresponding quantities c_ν and d_ν for ν. For any possible value of λ and μ denote by $T_{\mu,\lambda}$ a random variable with distribution $H_{\mu,\lambda}$. It is easily seen (as, for instance, in Theorem 2.1 of Möhle (2005)) that $S_{\mu,\lambda} := \dfrac{T_{\mu,\lambda} - \lambda c_\mu}{\sqrt{\lambda d_\mu}} \rightsquigarrow \Phi$ as $\lambda \to \infty$. Since the standard normal distribution is continuous, the distributions of $S_{\mu,\lambda}$ converge uniformly on $\mathbb{R}$ to Φ as $\lambda \to \infty$.

Let $1 < \kappa < 1/(2\rho)$. Then, since the distributions of $S_{\mu,\lambda}$ converge uniformly to Φ, there is λ_1 such that $1 - \Phi\left(\frac{U^\rho_{\mu,\lambda} - \lambda c_\mu}{\sqrt{\lambda d_\mu}}\right) < \kappa\rho$ for $\lambda > \lambda_1$ and $\lambda_2 > 0$ such that $\Phi\left(\frac{L^\rho_{\nu,\lambda+B} - (\lambda+B)c_\nu}{\sqrt{(\lambda+B)d_\nu}}\right) < \kappa\rho$ for all $\lambda > \lambda_2$. These two inequalities in turn imply that

$$\begin{aligned} U^\rho_{\mu,\lambda} &> \lambda c_\mu - \sqrt{\lambda d_\mu}\mathbf{z}_{\kappa\rho}, \\ L^\rho_{\nu,\lambda+B} &< (\lambda + B)c_\nu + \sqrt{(\lambda + B)d_\nu}\mathbf{z}_{\kappa\rho}. \end{aligned}$$

Since $c_\mu \geq c_\nu$ we can find λ_3 such that

$$(\lambda + B)c_\nu + \sqrt{(\lambda + B)d_\nu}\mathbf{z}_{\kappa\rho} < \lambda c_\mu - \sqrt{\lambda d_\mu}\mathbf{z}_{\kappa\rho} \quad \text{for all } \lambda \geq \lambda_3.$$

The first part of the result now follows by taking $\lambda_* := \lambda_1 \vee \lambda_2 \vee \lambda_3$. To prove the result for the measure ν_γ it suffices to see that we also have $\dfrac{T_{\nu_\gamma,\lambda+B} - (\lambda + B)c_{\nu_\gamma}}{\sqrt{(\lambda + B)d_{\nu_\gamma}}} \rightsquigarrow \Phi$, as $\lambda \to \infty$ (this is easily seen by analyzing the characteristic functions). The rest follows from the same argument used to prove the first part of the lemma. □

5.9.10 Proof of Lemma 5.4.4

Proof of (i): Let $s < t$. Note that $(Z_n)_{n=1}^\infty$ is a collection of i.i.d. random variables and $n\mathbb{P}_n(\zeta_0 + \frac{s}{n} < Z \leq \zeta_0 + \frac{t}{n})$ is permutation invariant, so the

Hewitt-Savage 0-1 law (see page 304 of Billingsley (1986)) implies that any convergent subsequence must converge to a constant. On the other hand, Lemma 5.3.3 implies that $n\mathbb{P}_n(\zeta_0 + \frac{s}{n} < Z \leq \zeta_0 + \frac{t}{n}) \rightsquigarrow \text{Poisson}((t-s)f(\zeta_0))$. Therefore, $\left(n\mathbb{P}_n(\zeta_0 + \frac{s}{n} < Z \leq \zeta_0 + \frac{t}{n})\right)_{n=1}^{\infty}$ has no almost surely convergent subsequence.

Proof of (ii)**:** Now, let $\delta \in (0, \frac{1}{4})$. From Proposition 5.3.2 we know that there is $B_\delta > 0$ such that $\mathbf{P}\left(n|\hat{\zeta}_n - \zeta_0| \leq B_\delta\right) > 1 - \delta$ for any $n \in \mathbb{N}$. Choose $h > 2B_\delta$ and take any increasing sequence of natural numbers n_k. Write $\hat{T}_k = n_k\mathbb{P}_{n_k}(\hat{\zeta}_{n_k} < Z \leq \hat{\zeta}_{n_k} + \frac{h}{n_k})$, $S_k = n_k\mathbb{P}_{n_k}(\zeta_0 - \frac{B_\delta}{n_k} < Z \leq \zeta_0 + \frac{h+B_\delta}{n_k})$ and $T_k = n_k\mathbb{P}_{n_k}(\zeta_0 + \frac{B_\delta}{n_k} < Z \leq \zeta_0 + \frac{h-B_\delta}{n_k})$. Then, $\left\{n_k|\hat{\zeta}_{n_k} - \zeta_0| \leq B_\delta\right\} \subset \left\{S_k \geq \hat{T}_k \geq T_k\right\}$ and therefore we have $\mathbf{P}\left(\hat{T}_k \geq T_k\right) \wedge \mathbf{P}\left(S_k \geq \hat{T}_k\right) > 1 - \delta$ for all k.

We know that $T_k \rightsquigarrow \text{Poisson}((h - 2B_\delta)f(\zeta_0))$ and $S_k \rightsquigarrow \text{Poisson}((h + 2B_\delta)f(\zeta_0))$, so in view of Lemma 5.9.4 with $B = 4B_\delta f(\zeta_0)$ and $\lambda = (h - 2B_\delta)f(\zeta_0)$, there is a number $h_* > 2B_\delta$ large enough so that whenever $h \geq h_*$ we can find two numbers $N_{1,h} < N_{2,h} \in \mathbb{N}$ with the property that, $\underline{\lim}_{k\to\infty} \mathbf{P}\left(T_k > N_{2,h}\right) > 2\delta$ and $\underline{\lim}_{k\to\infty} \mathbf{P}\left(S_k \leq N_{1,h}\right) > 2\delta$. Thus, for $h \geq h_*$, $\mathbf{P}\left(T_k > N_{2,h}\right) > 2\delta$ and $\mathbf{P}\left(S_k \leq N_{1,h}\right) > 2\delta$ for all but a finite number of k's. Therefore, for any k large enough, $\mathbf{P}\left(T_k > N_{2,h}\right) \wedge \mathbf{P}\left(S_k \leq N_{1,h}\right) > 2\delta$. Using the fact that $\mathbf{P}\left(S_k \geq \hat{T}_k \geq T_k\right) > 1 - \delta$ we get that $\mathbf{P}\left(\hat{T}_k \geq T_k > N_{2,h}\right) \wedge \mathbf{P}\left(N_{1,h} \geq S_k \geq \hat{T}_k\right) > \delta$ for all but finitely many k's. Thus, whenever $h \geq h_*$,

$$\mathbf{P}\left(\hat{T}_k \geq T_k > N_{2,h}, i.o.\right) > \delta \text{ and } \mathbf{P}\left(N_{1,h} \geq S_k \geq \hat{T}_k, i.o.\right) > \delta.$$

But for every $k \in \mathbb{N}$, the events $\left\{\hat{T}_k \geq T_k > N_{2,h}\right\}$ and $\left\{N_{1,h} \geq S_k \geq \hat{T}_k\right\}$ are permutation-invariant on the i.i.d. random vectors $X_1, \ldots, X_{n_k}$. Hence, the Hewitt-Savage 0-1 law implies that $\mathbf{P}\left(\hat{T}_k \geq T_k > N_{2,h}, i.o.\right) = 1$ and $\mathbf{P}\left(N_{1,h} \geq S_k \geq \hat{T}_k, i.o.\right) = 1$. Since $N_{1,h} < N_{2,h}$ it follows that $\hat{T}_k = n_k\mathbb{P}_{n_k}(\hat{\zeta}_{n_k} < Z \leq \hat{\zeta}_{n_k} + h/n_k)$ does not have an almost sure limit. But the choice of the subsequence n_k was arbitrary and independent of h_* so we can conclude that for any $h \geq h_*$, the sequence $\left\{n\mathbb{P}_n(\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{h}{n})\right\}_{n=1}^{\infty}$ does not converge in probability. Proceeding analogously, we can prove the same for $\left\{n\mathbb{P}_n(\hat{\zeta}_n - \frac{h}{n} < Z \leq \hat{\zeta}_n)\right\}_{n=1}^{\infty}$.

Proof of (iii)**:** We introduce some notation, for any two Borel probability measures μ and ν on $\mathbb{R}$ we write $\mu\bigstar\nu$ for their convolution and for $\lambda > 0$ we write CPoisson(μ, λ) for the compound Poisson distribution with intensity λ and compounding distribution μ. Let μ_α and μ_β be, respectively, the distributions under $\mathbb{P}$ of $\phi(\epsilon + \alpha_0)$ and $\phi(\epsilon + \beta_0)$.

Observe that depending on whether $t < 0$, $s < 0 < t$ or $s > 0$ we have that $n\mathbb{P}_n(\phi(Y)\mathbf{1}_{\zeta_0 + \frac{s}{n} < Z \leq \zeta_0 + \frac{t}{n}})$ converges weakly to CPoisson$(\mu_\alpha, (t-s)f(\zeta_0))$, CPoisson$(\mu_\alpha, sf(\zeta_0))\bigstar$CPoisson$(\mu_\beta, tf(\zeta_0))$ or CPoisson$(\mu_\beta, (t-s)f(\zeta_0))$, respectively. This follows easily from convergence of the corresponding characteristic functions. Considering that $\{(Y_n, Z_n)\}_{n=1}^{\infty}$ is a collection of i.i.d. random vectors and that $n\mathbb{P}_n(\phi(Y)\mathbf{1}_{\zeta_0 + \frac{s}{n} < Z \leq \zeta_0 + \frac{t}{n}})$ is permutation invariant for $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ the same argument as in (i) applies here as well.

Proof of (iv): We keep the notation used in the proof of (iii). The argument here is quite similar to the one used to show (ii). Assume without loss of generality that $\phi \leq 0$.

Now, let $\delta \in \left(0, \frac{1}{4}\right)$ and $N \in \mathbb{N}$. From Proposition 5.3.2 we know that there is $B_\delta > 0$ such that $\mathbf{P}\left(n|\hat{\zeta}_n - \zeta_0| \leq B_\delta\right) > 1 - \delta$ for any $n \in \mathbb{N}$. Choose $h > 2B_\delta$ and take any increasing sequence of natural numbers n_k. Write $\hat{T}^{\phi}_{k,h} = n_k\mathbb{P}_{n_k}(\phi(Y)\mathbf{1}_{\hat{\zeta}_{n_k} < Z \leq \hat{\zeta}_{n_k} + \frac{h}{n_k}})$, $S^{\phi}_{k,h} = n_k\mathbb{P}_{n_k}(\phi(Y)\mathbf{1}_{\zeta_0 - \frac{B_\delta}{n_k} < Z \leq \zeta_0 + \frac{h + B_\delta}{n_k}})$ and $T^{\phi}_{k,h} = n_k\mathbb{P}_{n_k}(\phi(Y)\mathbf{1}_{\zeta_0 + \frac{B_\delta}{n_k} < Z \leq \zeta_0 + \frac{h - B_\delta}{n_k}})$. Then, $\left\{n_k|\hat{\zeta}_{n_k} - \zeta_0| \leq B_\delta\right\} \subset \left\{S^{\phi}_{k,h} \leq \hat{T}^{\phi}_{k,h} \leq T^{\phi}_{k,h}\right\}$ and therefore we have $\mathbf{P}\left(\hat{T}^{\phi}_{k,h} \leq T^{\phi}_{k,h}\right) \wedge \mathbf{P}\left(S^{\phi}_{k,h} \leq \hat{T}^{\phi}_{k,h}\right) > 1 - \delta$ for all k.

We know that $T^{\phi}_{k,h} \rightsquigarrow$ CPoisson$(\mu_\beta, (h - 2B_\delta)f(\zeta_0))$ and

$$S^{\phi}_{k} \rightsquigarrow \text{CPoisson}(\mu_\alpha, 2B_\delta f(\zeta_0))\bigstar\text{CPoisson}(\mu_\beta, (h + B_\delta)f(\zeta_0))$$

$$\equiv \text{CPoisson}\left(\frac{B_\delta}{h+2B_\delta}\mu_\alpha + \frac{h+B_\delta}{h+2B_\delta}\mu_\beta, (h+2B_\delta)f(\zeta_0)\right),$$

as $k \to \infty$.

An application of Lemma 5.9.5 with $\mu = \nu = \mu_\beta$, $\gamma = \mu_\alpha$, $B = 4B_\delta f(\zeta_0)$, $r = \frac{1}{4}$ and $\lambda = (h-2B_\delta)f(\zeta_0)$, shows the existence of an $h_* > 2B_\delta$ large enough so that whenever $h \geq h_*$ we can find two numbers $R_{1,h} > R_{2,h} \in \mathbb{N}$ with the property that $\underline{\lim}_{k\to\infty} \mathbf{P}\left(T^\phi_{k,h} < R_{2,h}\right) > 2\delta$ and $\underline{\lim}_{k\to\infty} \mathbf{P}\left(S^\phi_{k,h} \geq R_{1,h}\right) > 2\delta$. Thus, for $h \geq h_*$, $\mathbf{P}\left(T^\phi_{k,h} < R_{2,h}\right) > 2\delta$ and $\mathbf{P}\left(S^\phi_{k,h} \geq R_{1,h}\right) > 2\delta$ for all but a finite number of k's. Therefore, for any k large enough, $\mathbf{P}\left(T^\phi_{k,h} < R_{2,h}\right) \wedge \mathbf{P}\left(S^\phi_{k,h} \geq R_{1,h}\right) > 2\delta$. Using the fact that $\mathbf{P}\left(S^\phi_{k,h} \leq \hat{T}^\phi_{k,h} \leq T^\phi_{k,h}\right) > 1-\delta$ we get that $\mathbf{P}\left(\hat{T}^\phi_{k,h} \leq T^\phi_{k,h} < R_{2,h}\right) \wedge \mathbf{P}\left(R_{1,h} \leq S^\phi_{k,h} \leq \hat{T}^\phi_{k,h}\right) > \delta$ for all but finitely many k's. Thus, whenever $h \geq h_*$,

$$\mathbf{P}\left(\hat{T}^\phi_{k,h} \leq T^\phi_{k,h} < R_{2,h}, i.o.\right) > \delta \text{ and } \mathbf{P}\left(R_{1,h} \leq S^\phi_{k,h} \leq \hat{T}^\phi_{k,h}, i.o.\right) > \delta.$$

The argument relying on the Hewitt-Savage 0-1 law applied in the proof of (ii) can be used to finish this proof.
A completely analogous proof applies for $\left\{n\mathbb{P}_n(\phi(Y)\mathbf{1}_{\hat{\zeta}_n - \frac{h}{n} < Z \leq \hat{\zeta}_n})\right\}_{n=1}^{\infty}$. $\square$

5.9.11 Proof of Lemma 5.4.6

We start by computing the characteristic functions of the weak limits of the last two components of the process $\tilde{\Xi}_n$ as defined in (5.17). Let $g_n(\xi)$ and $\psi_n(\xi)$ be the (unconditional) characteristic functions of $n\mathbb{P}^*_n(\mathbf{1}_{\zeta_0 < Z \leq \zeta_0 + \frac{t}{n}})$ and $n\mathbb{P}^*_n(\epsilon\mathbf{1}_{\zeta_0 < Z \leq \zeta_0 + \frac{t}{n}})$, respectively. Fix $\xi \in \mathbb{R}$ and write

$$\begin{aligned}
\Lambda_n &:= \mathbf{E}_{\mathfrak{X}}\left(\mathrm{e}^{i\xi n\mathbb{P}^*_n(\epsilon\mathbf{1}_{\zeta_0 < Z \leq \zeta_0 + \frac{t}{n}})}\right), \\
\Psi_n &:= n\mathbb{P}_n\left(\left(\mathrm{e}^{i\xi\epsilon} - 1\right)\mathbf{1}_{\zeta_0 < Z \leq \zeta_0 + \frac{t}{n}}\right), \\
\Psi^*_\xi &:= \sum_{1 \leq k \leq \nu(t)} \left(e^{i\xi\epsilon_k} - 1\right),
\end{aligned}$$

where $(\nu(s))_{s\geq 0}$ is a Poisson process with rate $f(\zeta_0)$ independent of $(\epsilon_n)_{n=1}^{\infty}$. Then, $\psi_n(\xi) = \mathbf{E}\left(\Lambda_n\right)$ and $|\Lambda_n| \leq 1$. By the conditional independence of the bootstrap samples, we have

$$\Lambda_n = \left(1 + \frac{1}{n}\Psi_n\right)^n.$$

We now consider the characteristic functions of the complex-valued random variables Ψ_n. Taking into account the independence of the X's, we obtain that for any $\eta \in \mathbb{R}^2$,

$$\begin{aligned}
\mathbf{E}\left(\mathrm{e}^{i\eta_1 \mathrm{Re}(\Psi_n) + i\eta_2 \mathrm{Im}(\Psi_n)}\right) &= \left(1 + \frac{1}{n}\mathbb{P}\left(\mathrm{e}^{i\eta_1(\cos(\xi\epsilon)-1)+i\eta_2\sin(\xi\epsilon)} - 1\right)\left(n\mathbb{P}(\mathbf{1}_{\zeta_0 < Z < \zeta_0 + \frac{t}{n}})\right)\right)^n \\
\mathbf{E}\left(\mathrm{e}^{i\eta_1 \mathrm{Re}(\Psi_n) + i\eta_2 \mathrm{Im}(\Psi_n)}\right) &\to \mathrm{e}^{tf(\zeta_0)\mathbf{E}\left(\mathrm{e}^{i\eta_1(\cos(\xi\epsilon)-1)+i\eta_2\sin(\xi\epsilon)}-1\right)} = \mathbf{E}\left(\mathrm{e}^{i\eta_1 \mathrm{Re}\Psi_\xi^* + i\eta_2 \mathrm{Im}\Psi_\xi^*}\right).
\end{aligned}$$

Therefore, $\Psi_n \rightsquigarrow \Psi_\xi^*$ and, from the continuous mapping theorem, $\Lambda_n \rightsquigarrow e^{\Psi_\xi^*}$. Thus, Lebesgue's Dominated Convergence Theorem implies

$$\psi_n(\xi) = \mathbf{E}\left(\Lambda_n\right) \to \mathbf{E}\left(\mathrm{e}^{\Psi_\xi^*}\right) = \mathrm{e}^{tf(\zeta_0)\left(\mathbf{E}\left(\mathrm{e}^{\mathrm{e}^{i\xi\epsilon}-1}\right)-1\right)} \quad \forall\, \xi \in \mathbb{R}. \tag{5.51}$$

With simpler arguments, we can also show that

$$g_n(\xi) \to \mathrm{e}^{tf(\zeta_0)\left(\mathrm{e}^{(\mathrm{e}^{i\xi\epsilon}-1)}-1\right)} \quad \forall\, \xi \in \mathbb{R}. \tag{5.52}$$

While (5.52) is immediately recognized as the characteristic function of a compound Poisson process with rate $f(\zeta_0)$ and compounding distribution Poisson(1), the characteristic function in (5.51) can be shown to correspond to another compound Poisson process which can be written as

$$\sum_{1\leq j\leq \nu(t)} \epsilon_j \tau_j, \tag{5.53}$$

where $(\tau_n)_{n=1}^{\infty} \overset{i.i.d.}{\sim}$ Poisson(1), $(\nu(s))_{s\geq 0}$ is a Poisson process with rate $f(\zeta_0)$, and $(\tau_n)_{n=1}^{\infty}$, $(\epsilon_n)_{n=1}^{\infty}$ and $(\nu(s))_{s\geq 0}$ are mutually independent.

Therefore, the fifth and sixth components of $\tilde{\Xi}_n$ as defined in (5.17) converge, respectively, to a compound Poisson process with rate $f(\zeta_0)$ and Poisson(1) as compounding distribution and to the process described in (5.53). A similar analysis shows the analogous results for the third and fourth components of $\tilde{\Xi}_n$. The first and second components of $\tilde{\Xi}_n$ can easily be seen (by using the Lindeberg-Feller Central Limit Theorem) to be asymptotically normal with mean 0 and variances $\sigma^2\mathbb{P}(Z \le \zeta_0)$ and $\sigma^2\mathbb{P}(Z > \zeta_0)$, respectively.

All these facts indicate that the finite dimensional distributions of the limiting process of $\tilde{\Xi}_n$ match those of the process $\tilde{\Xi}$. In fact, we can proceed as in the proof of Proposition 5.3.3 (i.e., proving tightness and convergence of the finite dimensional distributions using the Cramer-Wold device) to show (i) and (ii). For the sake of brevity, we omit the full technical details.

Then, arguing as in Proposition 5.3.2 one can show that the sequence $(\sqrt{n}(\alpha_n^* - \alpha_0), \sqrt{n}(\beta_n^* - \beta_0), n(\zeta_n^* - \zeta_0))$ is stochastically bounded and then conclude that the (unconditional) asymptotic distribution of $(\sqrt{n}(\alpha_n^* - \alpha_0), \sqrt{n}(\beta_n^* - \beta_0), n(\zeta_n^* - \zeta_0)$ is that of $\operatorname{sargmax}_{h\in\mathbb{R}^3}\{\tilde{E}^*(h)\}$, with $\tilde{E}^*(h)$ as defined in (5.18) and (5.19). For the sake of brevity we omit the full technical details of these arguments.

As $n(\zeta_n^* - \zeta_0) = n(\zeta_n^* - \hat{\zeta}_n) + n(\hat{\zeta}_n - \zeta_0)$, and if the ECDF bootstrap were consistent, the conditional distribution of $n(\zeta_n^* - \hat{\zeta}_n)$ (given the data) and the unconditional distribution of $n(\hat{\zeta}_n - \zeta_0)$ would have had the same weak limit. Then, as a consequence of Lemma 3.1 in Sen et al. (2010) (also see Theorem 2.2 in Kosorok (2008a)) the unconditional asymptotic distribution of $n(\zeta_n^* - \zeta_0)$ must be that of the sum of two independent copies of the asymptotic distribution of the $n(\hat{\zeta}_n - \zeta_0)$. The result now follows. $\square$

5.9.12 Proof of Lemma 5.4.7

Let $\mathbb{G}_n$ be the ECDF of $\epsilon_1, \ldots, \epsilon_n$. We first observe that

$$\int e^{i\xi x} d\mathbb{P}_n^{\epsilon}(x) = e^{-i\xi\bar{\epsilon}_n}\mathbb{P}_n\left(e^{i\xi\tilde{\epsilon}_n}\right)$$

and hence, for any $\xi \in \mathbb{R}$ with $|\xi| \leq \eta$ we have,

$$\begin{aligned} \left|\int e^{i\xi x} d\mathbb{P}_n^{\epsilon}(x) - e^{-i\xi\bar{\epsilon}_n}\int e^{i\xi x} d\mathbb{G}_n(x)\right| &= \left|\mathbb{P}_n\left(e^{i\xi\tilde{\epsilon}_n}\right) - \mathbb{P}_n\left(e^{i\xi\epsilon}\right)\right| \\ &\leq |\eta|\mathbb{P}_n\left(|\tilde{\epsilon}_n - \epsilon|\right) \end{aligned}$$

but $\mathbb{P}_n\left(|\tilde{\epsilon}_n - \epsilon|\right)$ is bounded from above by

$$|\hat{\alpha}_n - \alpha_0| + \left(|\alpha_0| + |\beta_0|\right)|\mathbb{P}_n(\mathbf{1}_{Z\leq\hat{\zeta}_n} - \mathbf{1}_{Z\leq\zeta_0})| + |\hat{\beta}_n - \beta_0|$$

which goes to zero almost surely as consequence of Lemmas 5.4.1 and 5.9.2 (iv), with $\mathbb{Q}_n = \mathbb{P}_n$. Thus,

$$\sup_{|\xi|\leq\eta}\left\{\left|\int e^{i\xi x} d\mathbb{P}_n^{\epsilon}(x) - e^{-i\xi\bar{\epsilon}_n}\int e^{i\xi x} d\mathbb{G}_n(x)\right|\right\} \xrightarrow{a.s.} 0$$

and (i) follows immediately because $\bar{\epsilon}_n = \mathbb{P}_n(\tilde{\epsilon}_n) \xrightarrow{a.s.} 0$ and $\mathbb{G}_n$ converges to G in total variation distance with probability one. The second assertion is seen to be true at once because G is assumed to be continuous and condition (i) implies that the characteristic functions of $\mathbb{P}_n^{\epsilon}$ converge to the characteristic function of G on the entire real line with probability one. Statements (ii) and (iii) are straightforward: On the one hand, we have shown that conditions (I)-(IV) hold for the ECDF, so Lemma 5.9.2 implies that $\int x^2 d\mathbb{P}_n^{\epsilon}(x) = \mathbb{P}_n(\tilde{\epsilon}_n^2) - \mathbb{P}_n(\tilde{\epsilon}_n)^2 \xrightarrow{a.s.} \sigma^2$. On the other hand,

$$\begin{aligned} \left|\int |x| d\mathbb{P}_n^{\epsilon} - \int |\epsilon| d\mathbb{P}_n\right| &= |\mathbb{P}_n(|\tilde{\epsilon}_n - \bar{\epsilon}_n| - |\epsilon|)| \\ &\leq \mathbb{P}_n(|\tilde{\epsilon}_n - \epsilon|) + |\bar{\epsilon}_n| \xrightarrow{a.s.} 0. \end{aligned}$$

To prove (iv), we first notice that

$$\int |x|^3 d\mathbb{P}_n^{\epsilon}(x) \leq |\bar{\epsilon}_n|^3 + 3|\bar{\epsilon}_n|^2\mathbb{P}_n\left(|\tilde{\epsilon}_n|\right) + 3|\bar{\epsilon}_n|\mathbb{P}_n\left(\tilde{\epsilon}_n^2\right) + \mathbb{P}_n\left(|\tilde{\epsilon}_n|^3\right).$$

Then, from Lemma 5.9.3 all but the last summand on the right-hand side converge almost surely. Hence, it suffices to show that $\overline{\lim}\,\mathbb{P}_n\left(|\tilde{\epsilon}_n|^3\right) < \infty$ w. p. 1. With this in mind, let $L_n = |\alpha_0| + |\hat{\alpha}_n| + |\beta_0| + |\hat{\beta}_n|$ and observe that

$$\mathbb{P}_n\left(|\tilde{\epsilon}_n|^3\right) \leq \mathbb{P}_n\left(|\epsilon|^3\right) + 3\mathbb{P}_n\left(|\epsilon|^2\right)L_n + 3\mathbb{P}_n\left(|\epsilon|\right)L_n^2 + L_n^3.$$

The result then is an immediate consequence of the third moment assumption on ϵ, the strong law of large numbers and the almost sure convergence of the least squares estimators. □

5.9.13 Proof of Proposition 5.4.2

Just as in the proof of Proposition 5.3.1 we have

$$\begin{aligned} -\frac{1}{n}\sum_{k=1}^{n}(\tilde{\epsilon}^*_{n,j})^2 &= R_n(\hat{\theta}_n) \\ &\leq R_n(\theta^*_n) \leq -\frac{1}{n}\sum_{j=1}^{n}(\tilde{\epsilon}^*_{n,j} + \hat{\alpha}_n - \alpha^*_n)^2\mathbf{1}_{Z_j<a} + (\tilde{\epsilon}^*_{n,j} + \hat{\beta}_n - \beta^*_n)^2\mathbf{1}_{Z_j>b} \end{aligned}$$

from which we can see that

$$|\gamma^*_n - \gamma_n|^2\mathbb{P}_n(Z<a)\wedge\mathbb{P}_n(Z>b) \leq$$
$$\frac{1}{n}\sum_{j=1}^{n}(\tilde{\epsilon}^*_{n,j})^2\mathbf{1}_{a\leq Z_j\leq b} + \frac{2}{n}|\gamma^*_n - \gamma_n|\left(\left|\sum_{j=1}^{n}\tilde{\epsilon}^*_{n,j}\mathbf{1}_{Z_j<a}\right| + \left|\sum_{j=1}^{n}\tilde{\epsilon}^*_{n,j}\mathbf{1}_{Z_j>b}\right|\right).$$

But the first of the terms on the right-hand side of the previous inequality is conditionally bounded in $\mathbb{L}_1$ (an upper bound for the conditional expectations is $\sup_{n\in\mathbb{N}}\left\{\int x^2 d\mathbb{P}^{\epsilon}_n(x)\right\} < \infty$). The terms $\frac{1}{n}\sum_{j=1}^{n}\tilde{\epsilon}^*_{n,j}\mathbf{1}_{Z_j<a}$ and $\frac{1}{n}\sum_{j=1}^{n}\tilde{\epsilon}^*_{n,j}\mathbf{1}_{Z_j>b}$ both have zero conditional expectation and conditional variances equal to $\frac{1}{n}\mathbb{P}_n(Z<a)\int x^2 d\mathbb{P}^{\epsilon}_n(x)$ and $\frac{1}{n}\mathbb{P}_n(Z>b)\int x^2 d\mathbb{P}^{\epsilon}_n(x)$ respectively. So we have that

$$\left|\frac{1}{n}\sum_{j=1}^{n}\tilde{\epsilon}^*_{n,j}\mathbf{1}_{Z_j<a}\right| + \left|\frac{1}{n}\sum_{j=1}^{n}\tilde{\epsilon}^*_{n,j}\mathbf{1}_{Z_j>b}\right| \underset{a.s.}{\xrightarrow{\mathbf{P}_{\mathfrak{X}}}} 0.$$

Thus,

$$|\gamma_n^* - \hat{\gamma}_n| = O_{\mathbf{P}_{\mathfrak{X}}}(1) \text{ almost surely.} \tag{5.54}$$

Now, let $Z_{(k)}$ be the k-th order statistic from the sample $(Z_1, \ldots, Z_n)$ and r_k a number such that $Z_{(k)} = Z_{r_k}$. For any $\zeta \in [a, b]$ define $m_\zeta = \max\{1 \leq j \leq n : Z_{(j)} \leq \zeta \wedge \hat{\zeta}_n\}$ and observe that we have

$$\frac{1}{n}\sum_{j=1}^{n} \tilde{\epsilon}_{n,j}^* \mathbf{1}_{Z_j \leq \zeta \wedge \hat{\zeta}_n} = \frac{1}{n}\sum_{1 \leq j \leq m_\zeta} \tilde{\epsilon}_{n,r_j}^*, \tag{5.55}$$

and thus

$$\sup_{\zeta \in [a,b]} \left\{ \left| \frac{1}{n}\sum_{j=1}^{n} \tilde{\epsilon}_{n,j}^* \mathbf{1}_{Z_j \leq \zeta \wedge \hat{\zeta}_n} \right| \right\} \leq \max_{1 \leq k \leq n} \left\{ \frac{1}{n} \left| \sum_{1 \leq j \leq k} \tilde{\epsilon}_{n,r_j}^* \right| \right\}. \tag{5.56}$$

But the indexes r_k and the order statistics are functions of $Z_1, \ldots, Z_n$ and therefore $\mathfrak{X}$-measurable. Hence, conditionally, $\sum_{1 \leq j \leq k} \tilde{\epsilon}_{n,r_j}^* \mathbf{1}_{Z_{r_j} \leq \zeta \wedge \hat{\zeta}_n}$ is a square integrable martingale with zero expectation. Hence, from Doob's submartingale inequality (see Williams (1991), Theorem 14.6, page 137) we get

$$\mathbf{P}_{\mathfrak{X}} \left(\max_{1 \leq k \leq n} \left\{ \frac{1}{n} \left| \sum_{1 \leq j \leq k} \tilde{\epsilon}_{n,r_j}^* \right| \right\} > \rho \right) \leq \frac{1}{n\rho^2} \mathbb{P}_n(\tilde{\epsilon}_n^2)$$

and consequently, equations (5.55) and (5.56) show that

$$\mathbf{P}_{\mathfrak{X}} \left(\left\| \frac{1}{n}\sum_{j=1}^{n} \tilde{\epsilon}_{n,j}^* \mathbf{1}_{Z_j \leq (\cdot) \wedge \hat{\zeta}_n} \right\|_{[a,b]} > \rho \right) \leq \frac{1}{\rho^2 n} \mathbb{P}_n(\tilde{\epsilon}_n^2) \xrightarrow{a.s.} 0. \tag{5.57}$$

Similar arguments give that (5.57) is also true if we replace $\mathbf{1}_{Z_j \leq (\cdot) \wedge \hat{\zeta}_n}$ by any of $\mathbf{1}_{(\cdot) < Z_j \leq \hat{\zeta}_n}$, $\mathbf{1}_{\hat{\zeta}_n < Z_j \leq (\cdot)}$ or $\mathbf{1}_{Z_j > (\cdot) \vee \hat{\zeta}_n}$. Now, if we write R_n like

$$\begin{aligned} R_n(\theta) = -\mathbb{P}_n^*(\tilde{\epsilon}_n^2) - \frac{2}{n}(\hat{\alpha}_n - \alpha) \sum_{j=1}^n \tilde{\epsilon}_{n,j}^* \mathbf{1}_{Z_j \leq \zeta \wedge \hat{\zeta}_n} - (\hat{\alpha}_n - \alpha)^2 \mathbb{P}_n(\mathbf{1}_{Z \leq \zeta \wedge \hat{\zeta}_n}) \\ -\frac{2}{n}(\hat{\beta}_n - \alpha) \sum_{j=1}^n \tilde{\epsilon}_{n,j}^* \mathbf{1}_{\hat{\zeta}_n < Z \leq \zeta} - (\hat{\beta}_n - \alpha)^2 \mathbb{P}_n(\mathbf{1}_{\hat{\zeta}_n < Z \leq \zeta}) \\ -\frac{2}{n}(\hat{\alpha}_n - \beta) \sum_{j=1}^n \tilde{\epsilon}_{n,j}^* \mathbf{1}_{\zeta < Z \leq \hat{\zeta}_n} - (\hat{\alpha}_n - \beta)^2 \mathbb{P}_n(\mathbf{1}_{\zeta < Z \leq \hat{\zeta}_n}) \\ -\frac{2}{n}(\hat{\beta}_n - \beta) \sum_{j=1}^n \tilde{\epsilon}_{n,j}^* \mathbf{1}_{Z > \zeta \vee \hat{\zeta}_n} - (\hat{\beta}_n - \beta)^2 \mathbb{P}_n(\mathbf{1}_{Z > \zeta \vee \hat{\zeta}_n}) \end{aligned} \tag{5.58}$$

(ii) follows immediately from (5.57), applied for all the four possible types of indicator functions. Note that the four terms on the far right of all the rows in the previous display vanish when we subtract M_n from R_n. Lemma 5.4.1 shows that (ii) implies (i), while Corollary 3.2.3 (ii), page 287, of van der Vaart and Wellner (1996) together with (5.54) allows one to derive (iii) from (i) and (ii). □

5.9.14 Proof of Lemma 5.4.8

The proof is analogous to the proof of Lemma 5.4.5. We again consider the number $h_* > 0$ defined in the statement of Lemma 5.4.4 and take $K \subset \mathbb{R}^3$ to be any compact rectangle containing the point $(h_*, 0, 0)$. To prove the theorem it suffices to show that the sequence $(\hat{E}_n(h_1, 0, 0))_{n=1}^\infty$ does not have a weak limit in probability whenever $h_1 \geq h_*$ and $(0, 0, h_1) \in K$. But in view of Lemma 5.4.4 this is straightforward because the (conditional) characteristic function of $\hat{E}_n(h_1, 0, 0)$ is given by

$$\left(\int e^{i2(\hat{\alpha}_n - \hat{\beta}_n)\xi x - i\xi(\hat{\alpha}_n - \hat{\beta}_n)^2} d\mathbb{P}_n^\epsilon(x) \right)^{n\mathbb{P}_n(\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{h_1}{n})}.$$

and Lemma 5.4.7 and the strong consistency of the least squares estimator imply that

$$\int e^{i2(\hat{\alpha}_n-\hat{\beta}_n)\xi x - i\xi(\hat{\alpha}_n-\hat{\beta}_n)^2} d\mathbb{P}_n^{\epsilon}(x) \xrightarrow{a.s.} e^{-i\xi(\alpha_0-\beta_0)^2}\varphi\left(2(\alpha_0-\beta_0)\xi\right).$$

Thus, for all ξ in a neighborhood of the origin, this characteristic function will converge if and only if $n\mathbb{P}_n(\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{h_1}{n})$ converges. We know that this is not the case from Lemma 5.4.4. □

5.9.15 Proof of Proposition 5.5.1

We will show that conditions (I)-(V) in Section 5.3 hold w.p. 1 for the bootstrap measures arising in this scheme. Note that (IV) is a consequence of Lemma 5.4.1. That $\|\mathbb{Q}_n - \mathbb{P}\|_{\mathcal{F}} \xrightarrow{a.s.} 0$ follows immediately from the fact that $\|\hat{F}_n - F\|_\infty \xrightarrow{a.s.} 0$. Now, for any $g = y\psi \in \mathcal{G}$ with $\psi \in \mathcal{F}$, we have

$$\begin{aligned}
\mathbb{Q}_n(g) &= \hat{\alpha}_n\mathbb{Q}_n(\mathbf{1}_{Z\leq\hat{\zeta}_n}\psi) + \hat{\beta}_n\mathbb{Q}_n(\mathbf{1}_{Z>\hat{\zeta}_n}\psi),\\
\mathbb{P}(g) &= \alpha_0\mathbb{P}(\mathbf{1}_{Z\leq\zeta_0}\psi) + \beta_0\mathbb{P}(\mathbf{1}_{Z>\zeta_0}\psi),
\end{aligned}$$

from which we see that

$$\begin{aligned}
\|\mathbb{Q}_n - \mathbb{P}\|_{\mathcal{G}} \leq{}& \left(|\hat{\alpha}_n - \alpha_0| + \left|\hat{\beta}_n - \beta_0\right|\right) + (|\alpha_0| + |\beta_0|)\,\|\mathbb{Q}_n - \mathbb{P}\|_{\mathcal{F}}\\
&+ (|\alpha_0| + |\beta_0|)\int_{\mathbb{R}} |\mathbf{1}_{z\leq\hat{\zeta}_n} - \mathbf{1}_{z\leq\zeta_0}|\hat{f}_n(z)dz.
\end{aligned}$$

Lebesgue's dominated convergence theorem shows that the last integral goes almost surely to zero and the strong consistency of the least squares estimators and property (I) now yields $\|\mathbb{Q}_n - \mathbb{P}\|_{\mathcal{G}} \xrightarrow{a.s.} 0$. Finally, we can write any $h \in \mathcal{H}$ in the form $h = y^2\psi$ for some $\psi \in \mathcal{F}$. Using this representation we obtain,

$$\begin{aligned}
\mathbb{Q}_n(h) &= \hat{\alpha}_n^2\mathbb{Q}_n(\mathbf{1}_{Z\leq\hat{\zeta}_n}\psi) + \hat{\beta}_n^2\mathbb{Q}_n(\mathbf{1}_{Z>\hat{\zeta}_n}\psi) + \mathbb{P}_n^{\epsilon}(\tilde{\epsilon}_n^2)\mathbb{Q}_n(\psi),\\
\mathbb{P}(h) &= \alpha_0^2\mathbb{P}(\mathbf{1}_{Z\leq\zeta_0}\psi) + \beta_0^2\mathbb{P}(\mathbf{1}_{Z>\zeta_0}\psi) + \sigma^2\mathbb{P}(\psi),
\end{aligned}$$

and the triangle inequality then implies that

$$\|\mathbb{Q}_n - \mathbb{P}\|_{\mathcal{H}} \leq (|\hat{\alpha}_n^2 - \alpha_0^2| + |\hat{\beta}_n^2 - \beta_0^2|) + (\alpha_0^2 + \beta_0^2 + \sigma^2) \|\mathbb{Q}_n - \mathbb{P}\|_{\mathcal{F}}$$
$$+ |\mathbb{P}_n^{\epsilon}(\tilde{\epsilon}_n^2) - \mathbb{P}(\epsilon^2)| + (\alpha_0^2 + \beta_0^2) \int_{\mathbb{R}} |\mathbf{1}_{z \leq \hat{\zeta}_n} - \mathbf{1}_{z \leq \zeta_0}| \hat{f}_n(z) dz \xrightarrow{a.s.} 0.$$

It remains to show (V). Observe that (5.6) and (5.7) hold automatically because under $\mathbb{Q}_n$, $\tilde{\epsilon}_n$ and Z are independent. Hence, we only require to show that (5.5) holds w.p. 1. As (5.23) holds, we have

$$\inf_{\zeta \in [c,d]} \left\{ \hat{f}_n(\zeta) \right\} \xrightarrow{a.s.} \inf_{\zeta \in [c,d]} \{ f(\zeta) \} > 0.$$

The mean value theorem implies that for any $\zeta, \xi \in [c, d]$, there is $\vartheta \in [0, 1]$ such that $|\hat{F}_n(\zeta) - \hat{F}_n(\xi)| = |\xi - \zeta| \hat{f}_n(\zeta + \vartheta(\xi - \zeta))$. It follows that for $\eta > 0$ small enough,

$$\inf_{0 < |\zeta - \hat{\zeta}_n| < \delta^2} \left\{ \frac{1}{|\zeta - \hat{\zeta}_n|} |\hat{F}_n(\zeta) - \hat{F}_n(\hat{\zeta}_n)| \right\} \geq \inf_{\zeta \in [c,d]} \left\{ \hat{f}_n(\zeta) \right\} \quad \forall\, n \in \mathbb{N}$$

and consequently (V) holds w.p.1 for all $\delta < \eta$ for all large n. ☐

5.9.16 Proof of Proposition 5.5.2

We already know that conditions (I)-(V) hold w.p. 1. Condition (VII) holds automatically because Z and $\tilde{\epsilon}_n$ are independent under $\mathbb{Q}_n$ and $\mathbb{Q}_n(\tilde{\epsilon}_n) = 0$. Lemma 5.4.7 (v) implies that condition (VIII) holds a.s. It remains to prove (VI).

Write $I = [c, d]$ and consider the sequence of events $\{A_N\}_{N \in \mathbb{N}}$ given by

$$A_N = \left[\hat{\zeta}_n - \frac{\delta}{n}, \hat{\zeta}_n + \frac{\eta}{n} \in I, \text{ almost always}, \ \forall\, \delta, \eta \in (0, N) \right] \cap \left[\|\hat{f}_n - f\|_I \to 0 \right].$$

Fix $N \in \mathbb{N}$, let ψ be the function $\psi(x) = e^{i\xi x}$ for some $\xi \in \mathbb{R}$ or the function $\psi(x) = |x|^p$, $p = 1, 2$, and $\eta, \delta > 0$ be any positive real numbers smaller than N. Then,

$$m_n \mathbb{Q}_n(\psi(\tilde{\epsilon}_n) \mathbf{1}_{\zeta_n - \frac{\delta}{n} < Z \leq \zeta_n + \frac{\eta}{n}}) = n \mathbb{P}_n^{\epsilon}(\psi) \int_{\hat{\zeta}_n - \frac{\delta}{n}}^{\hat{\zeta}_n + \frac{\eta}{n}} \hat{f}_n(x) dx.$$

Lemma 5.4.7 implies that $\mathbb{P}_n^{\epsilon}(\psi) \xrightarrow{a.s.} \mathbb{P}(\psi(\epsilon))$. And, when A_N holds, we also have

$$n\left|\int_{\hat{\zeta}_n-\frac{\delta}{n}}^{\hat{\zeta}_n+\frac{\eta}{n}} \hat{f}_n(x)dx - \int_{\hat{\zeta}_n-\frac{\delta}{n}}^{\hat{\zeta}_n+\frac{\eta}{n}} f(x)dx\right| \leq 2N\left\|\hat{f}_n - f\right\|_{[c,d]} \to 0.$$

Hence, condition (VI) holds for all $0 < \delta, \eta < N$ on A_N. But the strong consistency of the least squares estimators and the conditions on $\hat{f}_n$ imply that each of these events have probability one. Therefore, $\mathbf{P}(\cap_{N\in\mathbb{N}} A_N) = 1$. Hence, condition (VI) holds w.p.1 and the result follows from an application of Proposition 5.3.3. □

5.9.17 Proof of Proposition 5.5.3

Since $\mathbb{Q}_n$ is just the ECDF, the validity of conditions (I)-(IV) follows from the result established for the regular ECDF bootstrap and Lemma 5.4.1. (VIII) is a consequence of the strong law of large numbers. It remains to show (V)-(VII).

We start with (VI). First observe that $m_n\mathbb{P}(\psi(\epsilon)\mathbf{1}_{\zeta_0-\frac{\delta}{m_n}<Z\leq\zeta_0+\frac{\eta}{m_n}}) \to (\delta+\eta)f(\zeta_0)\mathbb{P}(\psi(\epsilon))$. We will proceed as follows: we will first use this simple observation just made to show that the following equations are true,

$$m_n\left\|\mathbb{P}_n(\psi(\epsilon)\mathbf{1}_{\zeta_0-\frac{(\cdot)}{m_n}<Z\leq\zeta_0}) - (\cdot)\mathbb{P}(\psi(\epsilon))f(\zeta_0)\right\|_K \xrightarrow{\mathbf{P}} 0 \quad (5.59)$$

$$m_n\left\|\mathbb{P}_n(\psi(\epsilon)\mathbf{1}_{\zeta_0<Z\leq\zeta_0+\frac{(\cdot)}{m_n}}) - (\cdot)\mathbb{P}(\psi(\epsilon))f(\zeta_0)\right\|_K \xrightarrow{\mathbf{P}} 0 \quad (5.60)$$

$$m_n\left\|\mathbb{P}_n(\psi(\tilde{\epsilon}_n)\mathbf{1}_{\hat{\zeta}_n-\frac{(\cdot)}{m_n}<Z\leq\hat{\zeta}_n}) - \mathbb{P}_n(\psi(\epsilon)\mathbf{1}_{\zeta_0-\frac{(\cdot)}{m_n}<Z\leq\zeta_0})\right\|_K \xrightarrow{\mathbf{P}} 0 \quad (5.61)$$

$$m_n\left\|\mathbb{P}_n(\psi(\tilde{\epsilon}_n)\mathbf{1}_{\hat{\zeta}_n<Z\leq\hat{\zeta}_n+\frac{(\cdot)}{m_n}}) - \mathbb{P}_n(\psi(\epsilon)\mathbf{1}_{\zeta_0<Z\leq\zeta_0+\frac{(\cdot)}{m_n}})\right\|_K \xrightarrow{\mathbf{P}} 0 \quad (5.62)$$

for any compact interval $K \subset \mathbb{R}$. All these facts put together will give

$$m_n\left\|\mathbb{P}_n(\psi(\tilde{\epsilon}_n)\mathbf{1}_{\hat{\zeta}_n-\frac{(\cdot)}{m_n}<Z\leq\hat{\zeta}_n}) - (\cdot)\mathbb{P}(\psi(\epsilon))f(\zeta_0)\right\|_K \xrightarrow{\mathbf{P}} 0 \quad (5.63)$$

$$m_n\left\|\mathbb{P}_n(\psi(\tilde{\epsilon}_n)\mathbf{1}_{\hat{\zeta}_n<Z\leq\hat{\zeta}_n+\frac{(\cdot)}{m_n}}) - (\cdot)\mathbb{P}(\psi(\epsilon))f(\zeta_0)\right\|_K \xrightarrow{\mathbf{P}} 0 \quad (5.64)$$

for any compact interval $K \subset \mathbb{R}$. Having achieved this, we will be able to conclude that (VI) holds in probability. For if (5.63) and (5.64) are both true, we can take an increasing sequence of compacts $(K_n)_{n=1}^{\infty}$ whose union is $\mathbb{R}$ and then for any subsequence $(n_k)_{k=1}^{\infty}$ find a further subsequence $(n_{k_s})_{s=1}^{\infty}$ such that

$$\mathbf{P}\left(m_{n_{k_s}}\left\|\mathbb{P}_{n_{k_s}}(\psi(\tilde{\epsilon}_{n_{k_s}})\mathbf{1}_{\hat{\zeta}_{n_{k_s}}-\frac{(\cdot)}{m_{n_{k_s}}}<Z\leq\hat{\zeta}_{n_{k_s}}})-(\cdot)\mathbb{P}(\psi(\epsilon))f(\zeta_0)\right\|_{K_s}>\frac{1}{s}\right) \quad < \quad \frac{1}{s^2}$$
$$\mathbf{P}\left(m_{n_{k_s}}\left\|\mathbb{P}_{n_{k_s}}(\psi(\tilde{\epsilon}_{n_{k_s}})\mathbf{1}_{\hat{\zeta}_{n_{k_s}}<Z\leq\hat{\zeta}_{n_{k_s}}+\frac{(\cdot)}{m_{n_{k_s}}}})-(\cdot)\mathbb{P}(\psi(\epsilon))f(\zeta_0)\right\|_{K_s}>\frac{1}{s}\right) \quad < \quad \frac{1}{s^2}.$$

The Borel-Cantelli Lemma will then imply that (VI) holds almost surely for the subsequence $(n_{k_s})_{s=1}^{\infty}$. Therefore, it suffices to show (5.59), (5.60), (5.61) and (5.62).

First consider the case where $\psi(\cdot) = |\cdot|$ and a positive number $\eta > 0$. Let $t \in \mathbb{R}$ and write

$$r_n = n\mathbb{P}\left(e^{i\frac{m_n}{n}t|\epsilon|} - 1 - \frac{m_n}{n}t|\epsilon|\right)\mathbb{P}(\mathbf{1}_{\zeta_0<Z\leq\zeta_0+\frac{\eta}{m_n}}).$$

Then, $|r_n| \leq t^2\sigma^2\frac{m_n}{n}m_n\mathbb{P}(\mathbf{1}_{\zeta_0<Z\leq\zeta_0+\frac{\eta}{m_n}}) \to 0$. The characteristic function of $m_n\mathbb{P}_n(|\epsilon|\mathbf{1}_{\zeta_0<Z\leq+\frac{\eta}{m_n}})$ can be written as

$$\varphi_n(t) = \left(1 + i\frac{m_n}{n}t\mathbb{P}(|\epsilon|)\mathbb{P}(\mathbf{1}_{\zeta_0<Z\leq\zeta_0+\frac{\eta}{m_n}}) + \frac{r_n}{n}\right)^n \to e^{it\eta\mathbb{P}(|\epsilon|)f(\zeta_0)}$$

and therefore

$$m_n\mathbb{P}_n(|\epsilon|\mathbf{1}_{\zeta_0<Z\leq\zeta_0+\frac{\eta}{m_n}}) \xrightarrow{\mathbf{P}} \eta f(\zeta_0)\mathbb{P}(|\epsilon|).$$

But

$$\sup_{n\in\mathbb{N}}\left\{\mathbf{E}\left(m_n\left\|\mathbb{P}_n(|\epsilon|\mathbf{1}_{\zeta_0<Z\leq\zeta_0+\frac{(\cdot)}{m_n}})\right\|_{[0,\eta]}\right)\right\} < \infty$$

and hence the sequence of processes $\left(m_n\mathbb{P}_n(|\epsilon|\mathbf{1}_{\zeta_0<Z\leq\zeta_0+\frac{(\cdot)}{m_n}})\right)_{n=1}^{\infty}$ is tight in $\mathcal{D}_{[0,\eta]}$. It follows that

$$m_n\mathbb{P}_n(|\epsilon|\mathbf{1}_{\zeta_0<Z\leq\zeta_0+\frac{(\cdot)}{m_n}}) \rightsquigarrow (\cdot)f(\zeta_0)\mathbb{P}(|\epsilon|) \text{ in } \mathcal{D}_{[0,\eta]}$$

but since the limiting process is continuous and deterministic we actually obtain

$$\left\| m_n \mathbb{P}_n(|\epsilon| \mathbf{1}_{\zeta_0 < Z \leq \zeta_0 + \frac{\cdot}{m_n}}) - (\cdot) f(\zeta_0) \mathbb{P}(|\epsilon|) \right\|_{[0,\eta]} \xrightarrow{\mathbf{P}} 0. \tag{5.65}$$

And with similar arguments one can also prove that

$$\left\| m_n \mathbb{P}_n(|\epsilon| \mathbf{1}_{\zeta_0 - \frac{(\cdot)}{m_n} < Z \leq \zeta_0}) - (\cdot) f(\zeta_0) \mathbb{P}(|\epsilon|) \right\|_{[0,\eta]} \xrightarrow{\mathbf{P}} 0. \tag{5.66}$$

Pick a positive number $\eta > 0$. Taking into account that $\epsilon \mathbf{1}_{\zeta_0 < Z \leq \zeta_0 + \frac{\eta}{m_n}} = (y - \beta_0) \mathbf{1}_{\zeta_0 < Z \leq \zeta_0 + \frac{\eta}{m_n}}$ and the analogous result for $\tilde{\epsilon}_n$ with $\hat{\zeta}_n$ and $\hat{\beta}_n$ instead of ζ_0 and β_0 we see that

$$\begin{aligned}
& m_n \left\| \mathbb{P}_n(|\tilde{\epsilon}_n| \mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{(\cdot)}{m_n}}) - \mathbb{P}_n(|\epsilon| \mathbf{1}_{\zeta_0 < Z \leq \zeta_0 + \frac{(\cdot)}{m_n}}) \right\|_{[0,\eta]} \leq \\
& m_n \left\| \mathbb{P}_n \left(|Y - \hat{\beta}_n| \left(\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{(\cdot)}{m_n}} - \mathbf{1}_{\zeta_0 < Z \leq \zeta_0 + \frac{(\cdot)}{m_n}} \right) \right) \right\|_{[0,\eta]} + \\
& m_n \left\| \mathbb{P}_n \left(\left(|Y - \hat{\beta}_n| - |Y - \beta_0| \right) \mathbf{1}_{\zeta_0 < Z \leq \zeta_0 + \frac{(\cdot)}{m_n}} \right) \right\|_{[0,\eta]}
\end{aligned}$$

and consequently

$$\begin{aligned}
& m_n \left\| \mathbb{P}_n(|\tilde{\epsilon}_n| \mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{(\cdot)}{m_n}}) - \mathbb{P}_n(|\epsilon| \mathbf{1}_{\zeta_0 < Z \leq \zeta_0 + \frac{(\cdot)}{m_n}}) \right\|_{[0,\eta]} \leq \\
& m_n \left\| \mathbb{P}_n \left(|Y - \beta_0| \left(\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{(\cdot)}{m_n}} - \mathbf{1}_{\zeta_0 < Z \leq \zeta_0 + \frac{(\cdot)}{m_n}} \right) \right) \right\|_{[0,\eta]} + \\
& |\hat{\beta}_n - \beta_0| m_n \left\| \mathbb{P}_n \left(\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{(\cdot)}{m_n}} - \mathbf{1}_{\zeta_0 < Z \leq \zeta_0 + \frac{(\cdot)}{m_n}} \right) \right\|_{[0,\eta]} + \\
& \\
& |\hat{\beta}_n - \beta_0| m_n \left\| \mathbb{P}_n \left(\mathbf{1}_{\zeta_0 < Z \leq \zeta_0 + \frac{(\cdot)}{m_n}} \right) \right\|_{[0,\eta]} .
\end{aligned} \tag{5.67}$$

We will show that each of the terms on the right-hand side of (5.67) goes to zero in probability. Since $n(\hat{\zeta}_n - \zeta_0) = O_{\mathbf{P}}(1)$, we know that for any $\delta > 0$ there is $R_\delta > 0$ such that $\mathbf{P}\left(n|\hat{\zeta}_n - \zeta_0| > R_\delta \right) < \delta$. Then,

$$\mathbf{P}\left(m_n \left\| \mathbb{P}_n \left(|Y - \beta_0| \left(\mathbf{1}_{\hat{\zeta}_n < Z \leq \hat{\zeta}_n + \frac{(\cdot)}{m_n}} - \mathbf{1}_{\zeta_0 < Z \leq \zeta_0 + \frac{(\cdot)}{m_n}} \right) \right) \right\|_{[0,\eta]} > \delta \right) \leq \delta +$$

$$\mathbf{P}\left(m_n\mathbb{P}_n\left(|\epsilon|\mathbf{1}_{\zeta_0-\frac{R_\delta}{n}<Z\le\zeta_0}\right)>\frac{\delta}{3}\right)+$$
$$\mathbf{P}\left(m_n\left\|\mathbb{P}_n\left(|\epsilon|\mathbf{1}_{\zeta_0<Z\le\zeta_0+\frac{(\cdot)}{m_n}+\frac{R_\delta}{n}}\right)\right\|_{[0,\eta]}>\frac{\delta}{3}\right)+$$

$$\mathbf{P}\left(m_n|\alpha_0-\beta_0|\mathbb{P}_n\left(\mathbf{1}_{\zeta_0-\frac{R_\delta}{n}<Z\le\zeta_0}\right)>\frac{\delta}{3}\right)$$

but from equations (5.65) and (5.66), and the fact that $\frac{m_n}{n}\to 0$, we actually get that all the terms of the right-hand side are asymptotically smaller than $\frac{\delta}{3}$. Thus,

$$\varlimsup_{n\to\infty}\mathbf{P}\left(m_n\left\|\mathbb{P}_n\left(|Y-\beta_0|\left(\mathbf{1}_{\hat{\zeta}_n<Z\le\hat{\zeta}_n+\frac{(\cdot)}{m_n}}-\mathbf{1}_{\zeta_0<Z\le\zeta_0+\frac{(\cdot)}{m_n}}\right)\right)\right\|_{[0,\eta]}>\delta\right)<2\delta. \tag{5.68}$$

An argument similar in spirit to the one just employed gives

$$\varlimsup_{n\to\infty}\mathbf{P}\left(m_n\left\|\mathbb{P}_n\left(\left(\mathbf{1}_{\hat{\zeta}_n<Z\le\hat{\zeta}_n+\frac{(\cdot)}{m_n}}-\mathbf{1}_{\zeta_0<Z\le\zeta_0+\frac{(\cdot)}{m_n}}\right)\right)\right\|_{[0,\eta]}>\delta\right)<\delta \tag{5.69}$$

while equation (5.70), for $\xi=0$, and the strong consistency of the least squares estimator give

$$|\hat{\beta}_n-\beta_0|m_n\left\|\mathbb{P}_n\left(\mathbf{1}_{\zeta_0<Z\le\zeta_0+\frac{(\cdot)}{m_n}}\right)\right\|_{[0,\eta]}\xrightarrow{\mathbf{P}}0.$$

Then, combining the last identity with (5.67), (5.68) and (5.69) we get

$$\lim_{\delta\to 0}\varlimsup_{n\to\infty}m_n\mathbf{P}\left(\left\|\mathbb{P}_n(|\tilde{\epsilon}_n|\mathbf{1}_{\hat{\zeta}_n<Z\le\hat{\zeta}_n+\frac{(\cdot)}{m_n}})-\mathbb{P}_n(|\epsilon|\mathbf{1}_{\zeta_0<Z\le\zeta_0+\frac{(\cdot)}{m_n}})\right\|_{[0,\eta]}>\delta\right)=0.$$

Completely analogous arguments prove that

$$m_n\left\|\mathbb{P}_n(|\tilde{\epsilon}_n|\mathbf{1}_{\hat{\zeta}_n-\frac{(\cdot)}{m_n}<Z\le\hat{\zeta}_n})-\mathbb{P}_n(|\epsilon|\mathbf{1}_{\zeta_0-\frac{(\cdot)}{m_n}<Z\le\zeta_0})\right\|_{[0,\eta]}\xrightarrow{\mathbf{P}}0.$$

Since $\eta>0$ was arbitrarily chosen, we have shown (IV) for $\psi(\cdot)=|\cdot|$. The case $\psi=|\cdot|^2$ is proven in a very similar manner. For the sake of brevity, we omit the proof.

Now, we consider the case where $\psi(x) = e^{i\xi x}$ for some $\xi \in \mathbb{R}$. Again, fix $\eta > 0$. We will proceed in the same way as before. Let $t \in \mathbb{R}$ and write

$$\rho_n = n\mathbb{P}\left(e^{i\frac{m_n}{n}t\cos(\xi\epsilon)} - 1 - \frac{m_n}{n}t\cos(\xi\epsilon)\right)\mathbb{P}(\mathbf{1}_{\zeta_0<Z\le\zeta_0+\frac{\eta}{m_n}}).$$

Then, $|\rho_n| \le t^2\frac{m_n}{n}m_n\mathbb{P}(\mathbf{1}_{\zeta_0<Z\le\zeta_0+\frac{\eta}{m_n}}) \to 0$. The characteristic function of $m_n\mathbb{P}_n(\cos(\xi\epsilon)\,\mathbf{1}_{\zeta_0<Z\le+\frac{\eta}{m_n}})$ can be written as

$$\varphi_n(t) = \left(1 + i\frac{m_n}{n}t\mathbb{P}(\cos(\xi\epsilon))\mathbb{P}(\mathbf{1}_{\zeta_0<Z\le\zeta_0+\frac{\eta}{m_n}}) + \frac{r_n}{n}\right)^n \to e^{it\eta\mathbb{P}(\cos(\xi\epsilon))f(\zeta_0)}$$

and therefore

$$m_n\mathbb{P}_n(\cos(\xi\epsilon)\,\mathbf{1}_{\zeta_0<Z\le\zeta_0+\frac{\eta}{m_n}}) \xrightarrow{\mathbf{P}} \eta f(\zeta_0)\mathbb{P}(\cos(\xi\epsilon)).$$

Applying the same arguments to the function $\sin(\xi\epsilon)$ we obtain that

$$m_n\mathbb{P}_n(\sin(\xi\epsilon)\,\mathbf{1}_{\zeta_0<Z\le\zeta_0+\frac{\eta}{m_n}}) \xrightarrow{\mathbf{P}} \eta f(\zeta_0)\mathbb{P}(\sin(\xi\epsilon)).$$

and hence

$$m_n\mathbb{P}_n(e^{i\xi\epsilon}\mathbf{1}_{\zeta_0<Z\le\zeta_0+\frac{\eta}{m_n}}) \xrightarrow{\mathbf{P}} \eta f(\zeta_0)\varphi\xi = \eta f(\zeta_0)\mathbb{P}(e^{i\xi\epsilon}).$$

The same tightness argument that was applied to prove (5.65) can be used here to conclude that

$$\left\|m_n\mathbb{P}_n(e^{i\xi\epsilon}\mathbf{1}_{\zeta_0<Z\le\zeta_0+\frac{\cdot}{m_n}}) - (\cdot)f(\zeta_0)\mathbb{P}(e^{i\xi\epsilon})\right\|_{[0,\eta]} \xrightarrow{\mathbf{P}} 0 \tag{5.70}$$

and similarly

$$\left\|m_n\mathbb{P}_n(e^{i\xi\epsilon}\mathbf{1}_{\zeta_0-\frac{(\cdot)}{m_n}<Z\le\zeta_0}) - (\cdot)f(\zeta_0)\mathbb{P}(e^{i\xi\epsilon})\right\|_{[0,\eta]} \xrightarrow{\mathbf{P}} 0. \tag{5.71}$$

Using the triangular inequality together with the definition of $\tilde{\epsilon}_n$ we get

$$m_n\left\|\mathbb{P}_n(e^{i\xi\tilde{\epsilon}_n}\mathbf{1}_{\hat{\zeta}_n<Z\le\hat{\zeta}_n+\frac{(\cdot)}{m_n}}) - \mathbb{P}_n(e^{i\xi\epsilon}\mathbf{1}_{\zeta_0<Z\le\zeta_0+\frac{(\cdot)}{m_n}})\right\|_{[0,\eta]} \le$$

$$m_n \left\| \mathbb{P}_n \left(\mathbf{1}_{\hat{\zeta}_n < Z \le \hat{\zeta}_n + \frac{(\cdot)}{m_n}} - \mathbf{1}_{\zeta_0 < Z \le \zeta_0 + \frac{(\cdot)}{m_n}} \right) \right\|_{[0,\eta]} +$$
$$m_n \left\| \mathbb{P}_n \left((e^{i\xi(Y-\hat{\beta}_n)} - e^{i\xi(Y-\beta_0)}) \mathbf{1}_{\zeta_0 < Z \le \zeta_0 + \frac{(\cdot)}{m_n}} \right) \right\|_{[0,\eta]} .$$

But (5.66) implies that

$$m_n \left\| \mathbb{P}_n \left(\mathbf{1}_{\hat{\zeta}_n < Z \le \hat{\zeta}_n + \frac{(\cdot)}{m_n}} - \mathbf{1}_{\zeta_0 < Z \le \zeta_0 + \frac{(\cdot)}{m_n}} \right) \right\|_{[0,\eta]} \xrightarrow{\mathbf{P}} 0$$

while (5.70) applied when $\xi = 0$ and the strong consistency of $\hat{\beta}_n$ yield

$$m_n \left\| \mathbb{P}_n \left((e^{i\xi(Y-\hat{\beta}_n)} - e^{i\xi(Y-\beta_0)}) \mathbf{1}_{\zeta_0 < Z \le \zeta_0 + \frac{(\cdot)}{m_n}} \right) \right\|_{[0,\eta]} \le$$
$$|\hat{\beta}_n - \beta_0| m_n \left\| \mathbb{P}_n \left(\mathbf{1}_{\zeta_0 < Z \le \zeta_0 + \frac{(\cdot)}{m_n}} \right) \right\|_{[0,\eta]} \xrightarrow{\mathbf{P}} 0.$$

Therefore,

$$m_n \left\| \mathbb{P}_n (e^{i\xi\tilde{\epsilon}_n} \mathbf{1}_{\hat{\zeta}_n < Z \le \hat{\zeta}_n + \frac{(\cdot)}{m_n}}) - \mathbb{P}_n (e^{i\xi\epsilon} \mathbf{1}_{\zeta_0 < Z \le \zeta_0 + \frac{(\cdot)}{m_n}}) \right\|_{[0,\eta]} \xrightarrow{\mathbf{P}} 0$$

which together with (5.70) proves that

$$\left\| m_n \mathbb{P}_n (e^{i\xi\tilde{\epsilon}_n} \mathbf{1}_{\hat{\zeta}_n < Z \le \hat{\zeta}_n + \frac{(\cdot)}{m_n}}) - (\cdot) f(\zeta_0) \mathbb{P}(e^{i\xi\epsilon}) \right\|_{[0,\eta]} \xrightarrow{\mathbf{P}} 0.$$

With completely analogous arguments one shows

$$m_n \left\| \mathbb{P}_n (e^{i\xi\tilde{\epsilon}_n} \mathbf{1}_{\hat{\zeta}_n - \frac{(\cdot)}{m_n} < Z \le \hat{\zeta}_n}) - (\cdot) f(\zeta_0) \mathbb{P}(e^{i\xi\epsilon}) \right\|_{[0,\eta]} \xrightarrow{\mathbf{P}} 0.$$

This proves that (VI) holds in probability.

We now proceed to prove that (V) and (VII) hold in probability. Before embarking in this task, we want to make the following remark. Consider that class of functions $\mathcal{C} := \{\epsilon \mathbf{1}_I(z) : I \subset \mathbb{R} \text{ is an interval}\}$. Then, this class has a square integrable envelope $|\epsilon|$ and $\mathbb{P}(\psi) = 0$ for any $\psi \in \mathcal{C}$. Therefore, the maximal inequality 3.1 from Kim and Pollard (1990) implies that $\|\mathbb{P}_n\|_{\mathcal{C}} = O_{\mathbf{P}}\left(n^{-\frac{1}{2}}\right)$. Similar observations also show that $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = O_{\mathbf{P}}\left(n^{-\frac{1}{2}}\right)$. All these considerations, in addition with Corollary 5.3.1, (5.61), (5.62), (5.59) and (5.60) show that

$$\sqrt{m_n}(\hat{\alpha}_n - \alpha_0) \xrightarrow{\mathbf{P}} 0 \tag{5.72}$$

$$\sqrt{m_n}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathbf{P}} 0 \tag{5.73}$$

$$m_n(\hat{\zeta}_n - \zeta_0) \xrightarrow{\mathbf{P}} 0 \tag{5.74}$$

$$\sqrt{m_n}\,\|\mathbb{P}_n\|_{\mathcal{C}} \xrightarrow{\mathbf{P}} 0 \tag{5.75}$$

$$\sqrt{m_n}\left\|\mathbb{P}_n(|\epsilon|\mathbf{1}_{\zeta_0-\frac{(\cdot)}{m_n}<Z\leq\zeta_0+\frac{(\cdot)}{m_n}})\right\|_K \xrightarrow{\mathbf{P}} 0 \tag{5.76}$$

$$\sqrt{m_n}\left\|\mathbb{P}_n(|\tilde{\epsilon}_n|\mathbf{1}_{\hat{\zeta}_n-\frac{(\cdot)}{m_n}<Z\leq\hat{\zeta}_n+\frac{(\cdot)}{m_n}})\right\|_K \xrightarrow{\mathbf{P}} 0 \tag{5.77}$$

$$\sqrt{m_n}\,\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \xrightarrow{\mathbf{P}} 0 \tag{5.78}$$

for any compact set $K \subset \mathbb{R}$.

Let $\eta > 0$ be fixed. Take any subsequence $(n_k)_{k=1}^{\infty}$ and find a further subsequence $(n_{k_s})_{s=1}^{\infty}$ such that all the statements in the previous display happen almost surely with the compact set K taken to be $K = [\zeta_0 - 2\eta, \zeta_0 + 2\eta]$. Now, for such a subsequence, there is $N \in \mathbb{N}$ such that $m_{n_{k_s}}|\zeta_0 - \hat{\zeta}_{n_{k_s}}| < \eta$ $\forall\ s \geq N$. Then, for any $\delta > 0$ and $s \geq N$, the following inequalities are true

$$\sup_{|\hat{\zeta}_{n_{k_s}}-\zeta|<\delta^2}\left\{|\mathbb{P}_{n_{k_s}}(\tilde{\epsilon}_{n_{k_s}}\mathbf{1}_{\zeta\wedge\hat{\zeta}_{n_{k_s}}<Z\leq\zeta\vee\hat{\zeta}_{n_{k_s}}})|\right\} \leq |\hat{\alpha}_{n_{k_s}} - \alpha_0| + |\hat{\beta}_{n_{k_s}} - \beta_0| +$$
$$\mathbb{P}_{n_{k_s}}(|\tilde{\epsilon}_{n_{k_s}}|\mathbf{1}_{\hat{\zeta}_{n_{k_s}}-\frac{\eta}{m_{n_{k_s}}}<Z\leq\hat{\zeta}_{n_{k_s}}+\frac{\eta}{m_{n_{k_s}}}}) + \mathbb{P}_{n_{k_s}}(|\epsilon|\mathbf{1}_{\zeta_0-\frac{2\eta}{m_{n_{k_s}}}<Z\leq\zeta_0+\frac{2\eta}{m_{n_{k_s}}}}) + \left\|\mathbb{P}_{n_{k_s}}\right\|_{\mathcal{C}}$$

$$\sup_{|\hat{\zeta}_{n_{k_s}}-\zeta|<\delta^2}\left\{|\mathbb{P}_{n_{k_s}}(\tilde{\epsilon}_{n_{k_s}}\mathbf{1}_{Z\leq\zeta\wedge\hat{\zeta}_{n_{k_s}}})| + |\mathbb{P}_{n_{k_s}}(\tilde{\epsilon}_{n_{k_s}}\mathbf{1}_{Z>\zeta\vee\hat{\zeta}_{n_{k_s}}})|\right\} \leq |\hat{\alpha}_{n_{k_s}} - \alpha_0| + |\hat{\beta}_{n_{k_s}} - \beta_0| +$$
$$\mathbb{P}_{n_{k_s}}(|\tilde{\epsilon}_{n_{k_s}}|\mathbf{1}_{\hat{\zeta}_{n_{k_s}}-\frac{\eta}{m_{n_{k_s}}}<Z\leq\hat{\zeta}_{n_{k_s}}+\frac{\eta}{m_{n_{k_s}}}}) + \mathbb{P}_{n_{k_s}}(|\epsilon|\mathbf{1}_{\zeta_0-\frac{2\eta}{m_{n_{k_s}}}<Z\leq\zeta_0+\frac{2\eta}{m_{n_{k_s}}}}) + \left\|\mathbb{P}_{n_{k_s}}\right\|_{\mathcal{C}}.$$

These last inequalities together with (5.72)-(5.78) imply that

$$\overline{\lim_{s\to\infty}}\sqrt{m_{n_{k_s}}}\sup_{|\hat{\zeta}_{n_{k_s}}-\zeta|<\delta^2}\left\{|\mathbb{P}_{n_{k_s}}(\tilde{\epsilon}_{n_{k_s}}\mathbf{1}_{\zeta\wedge\hat{\zeta}_{n_{k_s}}<Z\leq\zeta\vee\hat{\zeta}_{n_{k_s}}})|\right\} = 0\ a.s.$$

$$\overline{\lim_{s\to\infty}}\sqrt{m_{n_{k_s}}}\sup_{|\hat{\zeta}_{n_{k_s}}-\zeta|<\delta^2}\left\{|\mathbb{P}_{n_{k_s}}(\tilde{\epsilon}_{n_{k_s}}\mathbf{1}_{Z\leq\zeta\wedge\hat{\zeta}_{n_{k_s}}})| + |\mathbb{P}_{n_{k_s}}(\tilde{\epsilon}_{n_{k_s}}\mathbf{1}_{Z>\zeta\vee\hat{\zeta}_{n_{k_s}}})|\right\} = 0\ a.s.$$

The previous equations show that (5.6) and (5.7) in (V) as well as (VII) hold with probability one for the subsequence $(n_{k_s})_{s=1}^{\infty}$. We conclude by noting that if $\kappa = \inf_{z\in[a,b]} \{f(z)\}$, then the mean value theorem implies

$$\inf_{\frac{1}{\sqrt{m_{n_{k_s}}}}\leq|\zeta-\hat{\zeta}_{n_{k_s}}|<\delta^2} \left\{\frac{1}{|\zeta-\hat{\zeta}_{n_{k_s}}|}\mathbb{P}_{n_{k_s}}(\mathbf{1}_{\zeta\wedge\hat{\zeta}_{n_{k_s}}<Z\leq\zeta\vee\hat{\zeta}_{n_{k_s}}})\right\} \geq \kappa - \sqrt{m_{n_{k_s}}}\left\|\mathbb{P}_{n_{k_s}}-\mathbb{P}\right\|_{\mathcal{F}}$$

which in consequence shows

$$\varliminf_{s\to\infty} \inf_{\frac{1}{\sqrt{m_{n_{k_s}}}}\leq|\zeta-\hat{\zeta}_{n_{k_s}}|<\delta^2} \left\{\frac{1}{|\zeta-\hat{\zeta}_{n_{k_s}}|}\mathbb{P}_{n_{k_s}}(\mathbf{1}_{\zeta\wedge\hat{\zeta}_{n_{k_s}}<Z\leq\zeta\vee\hat{\zeta}_{n_{k_s}}})\right\} \geq \kappa > 0 \quad \text{a. s.}$$

This finishes the proof. $\square$

Bibliography

Abrevaya, J. and Huang, J. (2005). On the bootstrap of the maximum score estimator. *Econometrica*, 73:1175–1204.

Allon, G., Beenstock, M., Hackman, S., Passy, U., and Shapiro, A. (2007). Nonparametric estimation of concave production technologies by entropic methods. *J. Appl. Econometrics*, 4:795–816.

Attouch, B. and Wets, R. (1990). Epigraphical processes: laws of large numbers for random lsc functions. *Séminaire d'Analyse Convexe*, pages 13.1–13.29.

Aubin, J. and Frankowska, H. (2009). *Set-Valued Analysis*. Birkhäuser, Boston, MA, USA.

Banker, R. and Maindiratta, A. (1992). Maximum likelihood estimation of monotone and concave production frontiers. *J. Productiv. Anal.*, 3:401–415.

Beresteanu, A. (2007). Nonparametric estimation of regression functions under restrictions on partial derivatives. Available at: http://econ.duke.edu/ arie/shape.pdf (unpublished manuscript).

Bickel, P. J., Götze, F., and van Zwet, W. R. (1997). Resampling fewer than n observations: gains, losses, and remedies for losses. *Statist. Sinica*, 7:1–31.

Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley, New York, NY, USA.

Billingsley, P. (1986). *Probability and Measure*. John Wiley, New York, NY, USA.

Birke, M. and Dette, H. (2006). Estimating a convex function in nonparametric regression. *Scand. J. Statist.*, 34:384–404.

Boland, N. L. (1997). A dual-active-set algorithm for positive semi-definite quadratic programming. *Math. Program.*, 78:1–27.

Bose, A. and Chatterjee, S. (2001). Generalised bootstrap in non-regular m-estimation problems. *Stat. Prob. Lett.*, 55:319–328.

Bronšteĭn, E. M. (1976). ϵ-entropy of convex sets and functions. *Sibirsk. Mat. Zh.*, 17:508–514.

Bronšteĭn, E. M. (1978). Extremal convex functions. *Sibirsk. Mat. Zh.*, 19:10–18.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.*, 26:607–616.

Brunk, H. D. (1970). Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference*, pages 177–197. Camridge University Press, New York, NY, USA.

Chung, K. L. (2001). *A Course in Probability Theory*. Academic Press, San Diego, CA, USA.

Conway, J. (1985). *A Course in Functional Analysis*. Springer-Verlag, New York, NY, USA.

Courant, R. and John, F. (1999). *Introduction to Calculus and Analysis, Vol. II/1*. Springer, New York, NY, USA.

Cule, M. and Samworth, R. (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Stat.*, 4:254–270.

Cule, M., Samworth, R., and Stewart, M. (2010). Maximum likelihood estimation of a multidimensional log-concave density. *J. R. Stat. Soc. Ser. B (to appear)*.

Dempfle, A. and Stute, W. (2002). Nonparametric estimation of a discontinuity in regression. *Statist. Neerlandica*, 56:233–242.

Dudley, R. M. (1977). On second derivatives of convex functions. *Math. Scand.*, 41:159–174.

Dümbgen, L. (1991). The asymptotic behavior of some nonparametric change-point estimators. *Ann. Statist.*, 19:1471–1495.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38. SIAM, Philadelphia, USA.

Ethier, S. and Kurtz, T. (2005). *Markov Processes, Characterization and Convergence*. John Wiley & Sons, New York, NY, USA.

Ferger, D. (2004). A continuous mapping theorem for the argmax-functional in the non-unique case. *Statist. Neerlandica*, 48:83–96.

Folland, G. (1999). *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, New York, NY, USA.

Freedman, D. (1981). Bootstrapping regression models. *Ann. Statist.*, 9:1218–1228.

Gijbels, I., Hall, P., and Kneip, A. (1999). On the estimation of jump point in smooth curves. *Ann. Inst. Statist. Math.*, 51:231–251.

Gijbels, I., Hall, P., and Kneip, A. (2004). Interval and band estimation for curves with jumps. *J. Appl. Probab.*, 41:65–79.

Grenander, U. (1956). On the theory of mortality measurement. *Skan. Aktuarietidskr, Part II*, 39:125–153.

Groeneboom, P., Jongbloed, G., and Wellner, J. (2001). Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.*, 29:1653–1698.

Hanson, D. L. and Pledger, G. (1976). Consistency in concave regression. *Ann. Statist.*, 4:1038–1050.

Harville, D. (2008). *Matrix Algebra from a Statistician's Perspective*. Springer, New York, NY, USA.

Hazma, K. (1995). The smallest uniform upper bound on the distance between the mean and median of binomial and poisson distributions. *Statist. Probab. Lett.*, 23:21–25.

Hess, C. (1995). On the measurability of the conjugate and the subdifferential of a normal integrand. *J. of Convex Anal.*, 2:153–165.

Hildreth, C. (1954). Estimates of ordinates of concave functions. *J. Amer. Statist. Assoc.*, 49:598–619.

Hŭsková, M. and Kirch, C. (2008). Bootstrapping confidence intervals for the change-point of time series. *J. Time Ser. Anal.*, 29:947–972.

Johansen, S. (1974). The extremal convex functions. *Math. Scand.*, 41:61–68.

Kapoor, S. and Vaidya, P. M. (1986). Fast algorithms for convex quadratic programming and multicommodity flows. *Proceedings of the eighteenth annual ACM symposium on Theory of computing*, pages 147–159.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.*, 18:191–219.

Kosorok, M. (2008a). Bootstrapping the grenander estimator. *Beyond parametrics in interdisciplinary research: Festschrift in honor of professor Pranab K. Sen, IMS Collections*, 1:282–292.

Kosorok, M. (2008b). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, NY, USA.

Kosorok, M. and Song, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *Ann. Statist.*, 35:957–989.

Koul, H., Qian, L., and Surgailis, D. (2003). Asymptotics of m-estimators in two-phase linear regression models. *Stoc. Proc. Applic.*, 103:123–154.

Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. *Econom. J.*, 11:308–325.

Lan, Y., Banerjee, M., and Michailidis, G. (2009). Change-point estimation under adaptive sampling. *Ann. Statist.*, 37:1752–1791.

Lee, S. and Pun, M. (2006). On m out of n bootstrapping for nonstandard m-estimation with nuisance parameters. *J. Amer. Statist. Assoc.*, 101:1185–1197.

Lehmann, E. and Casella, G. (1998). *The Theory of Point Estimation*. Springer-Verlag, New York, USA.

Li, D. and Ling, S. (2012). On the least squares estimation of threshold autoregressive and moving-average models. *J. Econometrics*, 167:240–253.

Lim, E. and Glynn, P. (2012). Consistency in multi-dimensional convex regression. *Oper. Res.*, 60:196–208.

Luenberger, D. (1984). *Linear and Nonlinear Programming*. Addison-Wesley Publishing Company, Reading, MA, USA.

Maindiratta, A. and Sarath, B. (1997). On the consistency of maximum likelihood estimation of monotone and concave production frontiers. *J. Productiv. Anal.*, 8:239–246.

Mammen, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.*, 19:741–759.

Matzkin, R. L. (1991). Semiparametric estimation of monotone concave utility functions for polychotomous choice models. *Econometrica*, 59:1351–1327.

Matzkin, R. L. (1993). Nonparametric identification and estimation of polychotomous choice models. *J. Econometrics*, 58:137–168.

Mehrotra, S. and Sun, J. (1990). An algorithm for convex quadratic programming that requires $o(n^{3.5}l)$ arithmetic operations. *Math. Oper. Res.*, 15:342–363.

Möhle, M. (2005). Convergence results for compound poisson distributions and applications to the standard luria-delbrck distribution. *C. R. Acad. Sci. Paris*, 42:620–631.

Moreau, J. (1962). Decomposition orthogonal d'un espace hilbertien selon deux cones mutuellement polaires. *C. R. Acad. Sci. Paris*, 255:238–240.

Neuhaus, G. (1971). On weak convergence of stochastic processes with multidimensional time parameter. *Ann. Math. Statist.*, 42:1285–1295.

Nocedal, J. and Wright, S. (1999). *Numerical Optimization*. Springer, New York, NY, USA.

Pons, O. (2002). Estimation in a cox regression model with a change-point at an unknown time. *Statistics*, 36:101–124.

Pons, O. (2003). Estimation in a cox regression model with a change-point according to a threshold in a covariate. *Ann. Statist.*, 31:442–463.

Prakasa Rao, B. L. S. (1969). Estimation of a unimodal density. *Sankhya*, 31:23–36.

Rockafellar, T. R. (1969). Measurable dependence of convex sets and functions on parameters. *J. Math. Anal. Appl.*, 28:4–25.

Rockafellar, T. R. (1970). *Convex Analysis*. Princeton University Press, Princeton, NJ, USA.

Schuhmacher, D. and Dümbgen, L. (2010). Consistency of multivariate log-concave density estimators. *Statist. Probab. Lett.*, 80:376–380.

Schuhmacher, D., Hüsler, A., and Dümbgen, L. (2009). Multivariate log-concave distributions as a nearly parametric model. Tech. rep., University of Bern. Available at: http://arxiv.org/abs/0907.0250.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, New York, NY.

Seijo, E. and Sen, B. (2011a). Change-point in stochastic design regression and the bootstrap. *Ann. Statist.*, 39:1580–1607.

Seijo, E. and Sen, B. (2011b). Nonparametric least squares estimation of a multivariate convex regression function. *Ann. Statist.*, 39:1633–1657.

Sen, B., Banerjee, M., and Woodroofe, M. (2010). Inconsistency of bootstrap: The grenander estimator. *Ann. Statist.*, 38:1953–1977.

Seregin, A. and Wellner, J. (2009). Nonparametric estimation of multivariate convex-transformed densities. *Ann. Statist. (to appear)*. Available at: http://arxiv.org/abs/arXiv:0911.4151v1.

Silverman, B. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.*, 6:177–184.

Song, W. and Zhengjun, C. (2004). The generalized decomposition theorem in banach spaces and its applications. *J. Approx. Theory*, 129:167–181.

van der Vaart, A. W. (1998). *Asymptotics Statistics*. Cambridge University Press, Cambridge, UK.

van der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY, USA.

Varian, H. (1982a). The nonparametric approach to demand analysis. *Econometrica*, 50:945–973.

Varian, H. (1982b). The nonparametric approach to production analysis. *Econometrica*, 52:579–597.

Williams, D. (1991). *Probability with Martingales.* Cambridge University Press, Cambridge, UK.

Wu, C. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14:1261–1295.

Zhang, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.*, 30:528–555.

Appendix

Appendix A

Convex analysis

Lemma A.0.6 *Let $z \in \mathbb{R}^n$, $x_1, \ldots, x_n \in \mathbb{R}^d$ and define the function $g : \mathbb{R}^d \to \overline{\mathbb{R}}$ by*

$$g(x) = \inf \left\{ \sum_{k=1}^{n} \theta^k z^k : \sum_{k=1}^{n} \theta^k = 1,\ \sum_{k=1}^{n} \theta^k x_k = x,\ \theta \geq 0,\ \theta \in \mathbb{R}^n \right\}.$$

Then, g defines a convex function whose effective domain is $Conv\,(x_1, \ldots, x_n)$. Moreover, if $\mathcal{K}_{x,z}$ is the collection of all proper convex functions ψ such that $\psi(x_j) \leq z^j$ for all $j = 1, \ldots, n$, then $g = \sup_{\psi \in \mathcal{K}_{x,z}} \{\psi\}$.

Proof: To see that g defines a convex function, for any $x \in \mathbb{R}^d$ write

$$A_x = \left\{ \theta \in \mathbb{R}^n : \sum_{k=1}^{n} \theta^k = 1,\ \sum_{k=1}^{n} \theta^k x_k = x,\ \theta \geq 0 \right\}$$

and observe that for any $x, y \in \mathbb{R}^d$, $t \in (0,1)$, $\vartheta \in A_y$ and $\theta \in A_x$ we have $t\theta + (1-t)\vartheta \in A_{tx+(1-t)y}$ and hence

$$\frac{g\left(tx + (1-t)y\right) - (1-t)\sum_{k=1}^{n} \vartheta^k z^k}{t} \leq \sum_{k=1}^{n} \theta^k z^k.$$

Taking infimum over A_x and rearranging terms, we get

$$\frac{g\left(tx + (1-t)y\right) - tg(x)}{1-t} \leq \sum_{k=1}^{n} \vartheta^k z^k$$

and taking now the infimum over A_y gives the desired convexity. The convention that $\inf(\emptyset) = +\infty$ shows that the effective domain is precisely the convex hull of $x_1, \ldots, x_n$. Finally, for any $\psi \in \mathcal{K}_{x,z}$ and $x \in Conv\,(x_1, \ldots, x_n)$ we have, for $\theta \in \mathbb{R}^n$ with $\theta \geq 0$, $x = \sum_{j=1}^n \theta^j x_j$ and $\sum_{j=1}^n \theta^j = 1$,

$$\psi(x) \leq \sum_{j=1}^n \theta^j \psi(x_j) \leq \sum_{j=1}^n \theta^j z^j$$

since $\psi(x_j) \leq z^j$ for any $j = 1, \ldots, n$. The definition of g as an infimum then implies that $\psi(x) \leq g(x)\ \forall\ \psi \in \mathcal{K}_{x,z},\ x \in Conv\,(x_1, \ldots, x_n)$. The result then follows from the fact that $g \in \mathcal{K}_{x,z}$. $\square$

For the following results we use the notation introduced in Section 3.2, as they will concern α-monotone functions (see Definition 3.2.1).

Lemma A.0.7 *Let $\alpha \in \{-1, 0, 1\}^d$ and $f : \mathbb{R}^d \to \overline{\mathbb{R}}$. Then, f is α-monotone if and only if $f(x) \leq f(y)$ for all $x, y \in \mathbb{R}^d$ such that $(y - x) \in \mathbb{R}^d_\alpha$.*

Proof: If f is α-monotone and $(y - x) \in \mathbb{R}^d_\alpha$ then the equation $y = x + \sum_{j=1}^d \alpha^j(\alpha^j(y^j - x^j))\mathbf{e}_j$ implies that $f(x) \leq f(y)$. Conversely, if $f(x) \leq f(y)$ whenever $(y - x) \in \mathbb{R}^d_\alpha$, as $(x + r\alpha^j\mathbf{e}_j) - x = r\alpha^j\mathbf{e}_j \in \mathbb{R}^d_\alpha$ we can immediately conclude that $f(x) \leq f(x + r\alpha^j\mathbf{e}_j)$ for all $x \in \mathbb{R}^d$, $r \geq 0$ and $j = 1, \ldots, n$. $\square$

Lemma A.0.8 *Let $z \in \mathbb{R}^n$, $x_1, \ldots, x_n \in \mathbb{R}^d$, $\alpha \in \{-1, 0, 1\}^d$ and define the function $h_\alpha : \mathbb{R}^d \to \overline{\mathbb{R}}$ by*

$$h_\alpha(x) = \inf\left\{\sum_{k=1}^n \theta^k z^k : \sum_{k=1}^n \theta^k = 1,\ \vartheta + \sum_{k=1}^n \theta^k x_k = x,\ \theta \geq 0,\ \theta \in \mathbb{R}^n, \vartheta \in \mathbb{R}^d_{-\alpha}\right\}$$

Then, h_α defines a convex, α-monotone function whose effective domain is $Conv\,(x_1, \ldots, x_n) + \mathbb{R}^d_{-\alpha}$. Moreover, if $\mathcal{Q}^\alpha_{x,z}$ is the collection of all α-monotone, *proper convex functions* ψ *such that* $\psi(x_j) \leq z^j$ *for all* $j = 1, \ldots, n$, *then* $h_\alpha = \sup_{\psi \in \mathcal{Q}_{x,z}}\{\psi\}$.

Proof: The proof of the convexity of h_α is similar to the case of g in Lemma A.0.6. Now, if $x, y \in \mathbb{R}^d$ and $(y - x) \in \mathbb{R}^d_\alpha$, then for any $\theta \in \mathbb{R}^n$, $\vartheta \in \mathbb{R}^d_{-\alpha}$ with $\sum_{k=1}^n \theta^k = 1$, $\vartheta + \sum_{k=1}^n \theta^k x_k = y$, $\theta \geq 0$, we also have $\vartheta + (x-y) + \sum_{k=1}^n \theta^k x_k = x$ and $(\vartheta + (x-y)) \in \mathbb{R}^d_{-\alpha}$. Then, from the definition of h_α we see that $h_\alpha(x) \leq h_\alpha(y)$. Thus, h is α-monotone. That the effective domain of h_α is $Conv\,(x_1, \ldots, x_n) + \mathbb{R}^d_{-\alpha}$ is clear from the fact that for any x not belonging to that set, the infimum defining $h_\alpha(x)$ would be taken over the empty set. Finally, for any $\psi \in \mathcal{Q}^\alpha_{x,z}$ and $x \in Conv\,(x_1, \ldots, x_n) + \mathbb{R}^d_{-\alpha}$ we have, for $\theta \in \mathbb{R}^n$ and $\vartheta \in \mathbb{R}^d_{-\alpha}$ with $\theta \geq 0$, $x = \vartheta + \sum_{j=1}^n \theta^j x_j$ and $\sum_{j=1}^n \theta^j = 1$,

$$\psi(x) \leq \psi\left(\sum_{j=1}^n \theta^j x_j\right) \leq \sum_{j=1}^n \theta^j \psi(x_j) \leq \sum_{j=1}^n \theta^j z^j$$

since $\psi(x_j) \leq z^j$ for any $j = 1, \ldots, n$. The definition of h_α as an infimum then implies that $\psi(x) \leq h_\alpha(x)\ \forall\ \psi \in \mathcal{Q}^\alpha_{x,z}$, $x \in Conv\,(x_1, \ldots, x_n) + \mathbb{R}^d_{-\alpha}$. The result then follows from the fact that $h_\alpha \in \mathcal{Q}^\alpha_{x,z}$. $\square$

A.1 Polar coordinates based on boundaries of convex sets

Usual polar coordinates introduce a parametrization of $\mathbb{R}^d \setminus \{0\}$ based on the set $(0, \infty) \times \mathcal{S}^{d-1}$ where $\mathcal{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$ with respect to the Euclidian norm. This parametrization proves to be very useful for integration over spherical domains. Our aim in this section is to introduce a similar parametrization but now replacing $\mathcal{S}^{d-1}$ with the boundary of an arbitrary compact, convex set $\mathcal{X} \subset \mathbb{R}^d$ with nonempty interior. Throughout this section we will assume that $\mathcal{X}$ is a set of this type and that $x_0 \in \mathcal{X}^\circ$ is any point inside $\mathcal{X}$. We will use the notation $B(x,r)$ to denote the ball in $\mathbb{R}^d$ with radius r and center at x.

Lemma A.1.1 *For every $x \in \mathbb{R}^d \setminus \{0\}$ there is a unique $t_x > 0$ such that $x_0+t_x(x-x_0) \in \partial\mathcal{X}$, $x_0+t(x-x_0) \in \mathcal{X}^\circ$ for all $t \in (0,t_x)$ and $x_0+t(x-x_0) \in Ext(\mathcal{X})$ for all $t > t_x$.*

Proof: Without loss of generality we may assume that $x_0 = 0$. Consider the continuous function $\psi : \mathbb{R} \to \mathbb{R}^d$ given by $\psi(t) := tx$. Then, $\psi(0) \in \mathcal{X}^\circ$ and by compactness of $\mathcal{X}$ there is $M > 0$ such that $\psi(M) \in \mathbb{R}^d \setminus \mathcal{X}$. By the intermediate value theorem the set $\psi([0,M])$ must be connected in $\mathbb{R}^d$. It follows that there is $t_* \in (0,M)$ such that $\psi(t_*) \in \partial\mathcal{X}$. Now take $0 < t < t_*$. Since $0 \in \mathcal{X}^\circ$ there is $r > 0$ such that $B(0,r) \subset \mathcal{X}$. But then, $B(tx, \frac{t_*-t}{t_*}r) \subset Conv\left(\{t_* x\} \cup B(0,r)\right) \subset \mathcal{X}$ which in turn implies that $\psi(t) \in \mathcal{X}^\circ$. Finally, if there was a $t > t_*$ for which $\psi(t) \in \mathcal{X}$, we could switch the roles of t and t_* in the previous argument to see that we would have $\psi(t_*) \in \mathcal{X}^\circ$, a contradiction. This finishes the proof. $\square$

From the previous lemma we see that we can for every $x \in \mathbb{R}^d$ there are a unique $\rho_x := \frac{1}{t_x} > 0$ and $\xi_x \in \partial\mathcal{X}$ such that $x = x_0 + \rho_x \xi_x$. Consider now the function $\Phi_{\mathcal{X}} : \mathbb{R}^d \setminus \{x_0\} \to (0,\infty) \times \partial\mathcal{X}$ given by $\Phi_{\mathcal{X}}(x) := (\rho_x, \xi_x)$. Then we have the following result.

Lemma A.1.2 *Endow $\partial\mathcal{X}$ with the topology induced by the usual topology of $\mathbb{R}^d$ and $(0,\infty)\times\partial\mathcal{X}$ with the product topology. Then, $\Phi_{\mathcal{X}}$ is a homeomorphism.*

Proof: Assume, without loss of generality, that $x_0 = 0$. First, $\Phi_{\mathcal{X}}$ is clearly invertible with $\Phi_{\mathcal{X}}^{-1}(\rho,\xi) = \rho\xi$. The inequality $|\Phi_{\mathcal{X}}^{-1}(\rho_1,\xi_1) - \Phi_{\mathcal{X}}^{-1}(\rho_2,\xi_2)| \leq$

$|\rho_1 - \rho_2||\xi_1| + |\rho_2||\xi_1 - \xi_2|$ for any $\rho_1, \rho_2 > 0$, $\xi_1, \xi_2 \in \partial\mathcal{X}$ shows that $\Phi_{\mathcal{X}}^{-1}$ is continuous. On the other hand, for any $x \in \mathbb{R}^d \setminus \{0\}$ and $r > 0$ we have that $B\left((t_x \wedge 1)x, \frac{t_x \wedge 1}{t_x \vee 1} r\right) \subset Conv\left(\{0\}, B\left((t_x \vee 1)x, r\right)\right)$, where $t_x = \rho_x^{-1}$. The latter fact implies that ξ_x is a continuous function of x. Finally, the identity $\rho_x = \frac{|x|}{|\xi_x|}$ shows that ρ_x is also a continuous function of x. Hence, Φ_X is continuous with continuous inverse. $\square$

We will now present a generalization of the traditional change-of-variables formula for spherical coordinates. We will denote by λ_d the Lebesgue measure on $\mathbb{R}^d$ and by τ_d the measure on $(0, \infty)$ given by $\tau_d(dt) = t^{d-1}dt$.

Lemma A.1.3 (Change-of-variables Formula) *Consider the Borel measure $m_d^{\mathcal{X}}(\cdot) = \lambda_d \Phi_{\mathcal{X}}^{-1}(\cdot)$ on $(0,\infty) \times \partial\mathcal{X}$. Then, there is a unique Borel measure $\gamma_{\mathcal{X}}^{d-1}$ on $\partial\mathcal{X}$ such that $m_d^{\mathcal{X}} = \tau_d \times \gamma_{\mathcal{X}}$. Moreover, for any measurable function $f : \mathbb{R}^d \to \mathbb{C}$ which is either nonnegative or integrable (with respect to λ_d) we have:*

$$\int f(x)dx = \int f \circ \Phi_{\mathcal{X}}^{-1}(s,\xi)\tau_d(ds)\gamma_{\mathcal{X}}(d\xi) = \iint_{\mathbb{R}_+ \times \partial\mathcal{X}} f(x_0 + s\xi)s^{d-1}ds\gamma_{\mathcal{X}}(d\xi).$$

Proof: This result is a generalization of Theorem 2.49 in page 78 of Folland (1999). We refer the reader to the proof provided there. Although that result refers only to the case when $x_0 = 0$, $\mathcal{X} = B(0,1)$ and $\partial\mathcal{X} = \mathcal{S}^{d-1}$, all the arguments remain valid for arbitrary $\mathcal{X}$ and x_0. $\square$

The measure $\gamma_{\mathcal{X}}$ of the previous theorem can be thought as a "surface-area" measure on $\partial\mathcal{X}$. A more general version of this formula is known in the geometric measure theory literature as the co-area formula.

A.2 Restrictions of convex functions to compact, convex subsets of their effective domains

The following results turn out to be useful in the analysis of the local behavior of convex functions. For a convex function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ and a convex set $\mathcal{X} \subset \mathbb{R}^d$ we denote by $\mathcal{K}_{\mathcal{X},f}$ the class of all convex functions g such that $g(x) \leq f(x)$ for all $x \in \mathbb{R}^d \setminus \mathcal{X}^\circ$.

Lemma A.2.1 *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be a closed, proper convex function such that $dom(f)^\circ \neq \emptyset$ and $\mathcal{X} \subset dom(f)^\circ$ be a compact, convex set with nonempty interior. Consider the function $\underline{f} : \mathbb{R}^d \to \overline{\mathbb{R}}$ given by:*

$$\underline{f}(x) = \sup_{\substack{\xi \in \partial f(y) \\ y \in \mathbb{R}^d \setminus \mathcal{X}^\circ}} \{\langle \xi, x\rangle - f^*(\xi)\}.$$

Then,

(i) $\underline{f} \leq f$*; in particular* $\underline{f} \in \mathcal{K}_{\mathcal{X},f}$*.*

(ii) $\underline{f}(x) = f(x)$ *for every* $x \in \mathbb{R}^d \setminus \mathcal{X}^\circ$*.*

(iii) For $x \in \mathcal{X}$ *we have*

$$\underline{f}(x) = \sup_{\substack{\xi \in \partial f(y) \\ y \in \partial\mathcal{X}}} \{\langle \xi, x\rangle - f^*(\xi)\}.$$

(iv) If $x \in \mathcal{X}^\circ$ *then,*

$$f(x) - \underline{f}(x) = \inf_{\substack{\xi \in \partial f(y) \\ y \in \partial\mathcal{X}}} \{f(x) - f(y) - \langle \xi, x - y\rangle\}.$$

Proof: Since f is closed, Theorem 12.2, page 104 of Rockafellar (1970) implies that $f = f^{**}$. Thus, from Corollary 12.2.2, page 104 in the same reference we have $f(x) = \sup_{\xi \in ri(dom(f^*))} \{\langle \xi, x\rangle - f^*(x)\}$, where $ri(dom(f^*))$ denotes the *relative interior* of $dom(f^*)$. But the remarks in page 227 of Rockafellar (1970) imply that

$$ri(dom(f^*)) \subset \bigcup_{x \in \mathbb{R}^d} \partial f(x) \subset dom(f^*) \subset \mathbb{R}^d.$$

Therefore we get the following identity

$$\sup_{\xi \in dom(f^*)} \{\langle \xi, x\rangle - f^*(\xi)\} = f(x) = \sup_{\substack{\xi \in \partial f(y) \\ y \in \mathbb{R}^d}} \{\langle \xi, x\rangle - f^*(\xi)\}.$$

It follows immediately that $\underline{f} \leq f$. Now, let $x \in \mathbb{R}^d \setminus \mathcal{X}^\circ$. Choose $y \in \mathcal{X}^\circ$ and consider the one-dimensional convex function $f_y(t) = f(x + t(y - x))$. Note that $\partial f_y(t) = (y - x)'\partial f(x + t(y - x))$ as a consequence of Theorem 23.9, page 225 in Rockafellar (1970). From Lemma A.1.1 there is $0 < t_* < 1$ such that $y_* := x + t_*(y - x) \in \partial\mathcal{X}$. Choose $\xi_* \in \partial f(y_*)$ and $\xi \in \partial f(y)$. Note that $\langle \xi_*, y - x\rangle \in \partial f_y(t_*)$ and $\langle \xi, y - x\rangle \in \partial f_y(1)$ which implies that $\langle \xi_*, y - x\rangle \leq \langle \xi, y - x\rangle$ as $0 < t_* < 1$. Thus, using that $\xi \in \partial f(y)$ (so $f(y) + f^*(\xi) = \langle \xi, y\rangle$ by Theorem 23.5 in page 218 of Rockafellar (1970)) we have

$$\langle \xi, x\rangle - f^*(\xi) = \langle \xi, x - y\rangle + f(y) \leq f(y_*) - t_*\langle \xi, x - y\rangle \leq f(y_*) - t_*\langle \xi_*, x - y\rangle = \langle \xi_*, x\rangle - f^*(\xi_*).$$

We have thus shown that for any $x \in \mathbb{R}^d \setminus \mathcal{X}^\circ$, $y \in \mathcal{X}^\circ$ and $\xi \in \partial f(y)$ there are $y_* \in \partial\mathcal{X}$ and $\xi_* \in \partial f(y_*)$ such that $\langle \xi, x\rangle - f^*(\xi) \leq \langle \xi_*, x\rangle - f^*(\xi_*)$. It follows that

$$f(x) = \sup_{\substack{\xi \in \partial f(y) \\ y \in \mathbb{R}^d}} \{\langle \xi, x\rangle - f^*(\xi)\} = \sup_{\substack{\xi \in \partial f(y) \\ y \in \mathbb{R}^d \setminus \mathcal{X}^\circ}} \{\langle \xi, x\rangle - f^*(\xi)\} = \underline{f}(x) \quad \forall x \in \mathbb{R} \setminus \mathcal{X}^\circ,$$

which shows (ii). A similar argument can be used to prove that for any $x \in \mathcal{X}^\circ$, any $y \in \mathbb{R}^d \setminus \mathcal{X}^\circ$ and any $\xi \in \partial f(y)$ there are $\tilde{y} \in \partial\mathcal{X}$ and $\tilde{\xi} \in \partial f(\tilde{y})$ such that $\langle \xi, x\rangle - f^*(\xi) \le \langle \tilde{\xi}, x\rangle - f^*(\tilde{\xi})$. This implies (iii). The last statement follows from (iii) and the identity $f(y) + f^*(\xi) = \langle y, \xi\rangle$ for all $y \in \mathbb{R}^d$ and $\xi \in \partial f(y)$. □

Lemma A.2.2 *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be a closed, proper convex function such that $dom(f)^\circ \neq \emptyset$ and $\mathcal{X} \subset dom(f)^\circ$ be a compact, convex set with nonempty interior. Let $\overline{f} = \sup_{\psi \in \mathcal{K}_{\mathcal{X},f}}\{\psi\}$. Then,*

(i) $f \le \overline{f}$.

(ii) $\overline{f}(x) = f(x)$ *for all* $x \in \mathbb{R}^d \setminus \mathcal{X}^\circ$.

(iii) For every $x \in \mathcal{X}^\circ$ *we have:*

$$\overline{f}(x) = \min_{\substack{x_1,\dots,x_n \in \partial\mathcal{X},\ n\in\mathbb{N} \\ \theta^1,\dots,\theta^n \ge 0,\ \sum\theta^j = 1 \\ \sum\theta^j x_j = x}} \left\{ \sum_{j=1}^n \theta^j f(x_j) \right\}$$

(iv) For every $x \in \mathcal{X}^\circ$ *we have:*

$$\overline{f}(x) - f(x) = \min_{\substack{x_1,\dots,x_n \in \partial\mathcal{X},\ n\in\mathbb{N} \\ \theta^1,\dots,\theta^n \ge 0,\ \sum\theta^j = 1 \\ \sum\theta^j x_j = x}} \left\{ \sum_{j=1}^n \theta^j (f(x_j) - f(x)) \right\}.$$

(v) For every $x \in \mathcal{X}^\circ$ *and* $w \in \mathcal{X}$ *we have:*

$$\frac{1}{2}\inf_{\substack{\xi\in\partial f(x)\\ y\in\partial\mathcal{X}}}\{f(y)-f(x)-\langle\xi, y-x\rangle\} \le \overline{f}(x)-f(x) \le \sup_{\substack{\xi\in\partial f(w)\\ y\in\partial\mathcal{X}}}\{f(y)-f(w)-\langle\xi, y-w\rangle\}.$$

Proof: The first two statements are obvious consequences of the definition of the function. Now, let $F \subset \mathbb{R}^{d+1}$ be given by $F := \{(x,t) \in epi(f) : x \in$

$\mathbb{R}^d \setminus \mathcal{X}^\circ\}$ and consider the function

$$\tilde{f}(x) := \inf\{t : (x,t) \in conv(F)\} = \inf_{\substack{x_1,\ldots,x_n \in \mathbb{R}^d \setminus \mathcal{X}^\circ \\ \theta^1,\ldots,\theta^n \geq 0,\ \sum \theta^j = 1 \\ \sum \theta^j x_j = x}} \left\{ \sum_{j=1}^n \theta^j f(x_j) \right\}. \quad \text{(A.1)}$$

Note that the infimum in (A.1) is actually a minimum form Corollary 19.1.2 in page 173 of Rockafellar (1970). From Theorem 5.3, page 33 in Rockafellar (1970) we know that $\tilde{f}$ is a convex function. Note that $F \subset \{(x,t) \in epi(g) : x \in \mathbb{R}^d \setminus \mathcal{X}^\circ\}$ for every $g \in \mathcal{K}_{\mathcal{X},f}$. Hence, $conv(F) \subset epi(g)$ and $g \leq \tilde{f}$ for every $g \in \mathcal{K}_{\mathcal{X},f}$ and, consequently, $\tilde{f} \geq \overline{f}$. On the other hand, from the definition of $\tilde{f}$ it is obvious that $\tilde{f} \equiv f$ on $\mathbb{R}^d \setminus \mathcal{X}^\circ$. Thus, $\tilde{f} = \overline{f}$. To show (iii) it suffices to argue that for $x \in \mathbb{R}^d \setminus \mathcal{X}^\circ$ we can take the infimum in (A.1) with the x's ranging only on $\partial\mathcal{X}$. To achieve this, we will prove that for any $x \in \mathcal{X}^\circ$ and $x_1, \ldots, x_{n+1} \in \mathbb{R}^d \setminus \mathcal{X}^\circ$ with $x \in conv(x_1, \ldots, x_{n+1})$, $x_{n+1} \in Ext(\mathcal{X})$ and x expressed as the convex combination $x = \theta^1 x_1 + \cdots + \theta^{n+1} x_{n+1}$ there are $\tilde{x}_{n+1} \in \partial\mathcal{X}$ and nonnegative coefficients $\tilde{\theta}^1, \ldots, \tilde{\theta}^{n+1}$ such that x can be expressed as the convex combination $x = \tilde{\theta}^1 x_1 + \cdots + \tilde{\theta}^n x_n + \tilde{\theta}^{n+1} \tilde{x}_{n+1}$ and $\tilde{\theta}^1 f(x_1) + \cdots + \tilde{\theta}^n f(x_n) + \tilde{\theta}^{n+1} f(\tilde{x}_{n+1}) \leq \theta^1 f(x_1) + \cdots + \theta^{n+1} f(x_{n+1})$. From Lemma A.1.1, there is $0 < \tilde{t} < 1$ such that $\tilde{x}_{n+1} = x + \tilde{t}(x^{n+1} - x) \in \partial\mathcal{X}$. Let $\tilde{\theta}^k := \frac{\tilde{t}\theta^k}{\tilde{t} + (1-\tilde{t})\theta^{n+1}}$ for $k = 1, \ldots, n$ and $\tilde{\theta}^{n+1} := \frac{\theta^{n+1}}{\tilde{t} + (1-\tilde{t})\theta^{n+1}}$. Then it is easily seen that $\tilde{x}^{n+1}$ and $\tilde{\theta}^1, \ldots, \tilde{\theta}^{n+1}$ satisfy the desired condition.

It remains to show (v) as (iv) is an obvious consequence of (iii). For any $x_1, \ldots, x_n \in \partial\mathcal{X}$, any $\xi \in \partial f(x)$ and any $J \subset \{1, \ldots, n\}$, if x is the convex combination of $x_1, \ldots, x_n$ with coefficients $\theta^1, \ldots, \theta^n$ we have:

$$\begin{aligned}
\sum_{k=1}^n \theta^k f(x_k) - f(x) &\geq \sum_{k\in J} \theta^k (f(x_k) - f(x)) + \sum_{k \notin J} \theta^k \langle \xi, x_k - x \rangle \\
&\geq \sum_{k\in J} \theta^k (f(x_k) - f(x)) - \sum_{k \in J} \theta^k \langle \xi, x_k - x \rangle \\
&\geq \sum_{k\in J} \theta^k (f(x_k) - f(x) - \langle \xi, x_k - x \rangle).
\end{aligned}$$

Applying the same argument to the complement of J and taking infimum over all elements $x_k \in \partial\mathcal{X}$ and $\xi \in \partial f(x)$ we obtain:

$$\begin{aligned}\sum_{k=1}^{n}\theta^k f(x_k) - f(x) &\geq \left(\sum_{k\in J}\theta^k\right) \vee \left(\sum_{k\notin J}\theta^k\right) \inf_{\substack{\eta\in\partial f(x)\\ y\in\partial\mathcal{X}}}\{f(y) - f(x) - \langle\eta, y-x\rangle\} \\ &\geq \frac{1}{2}\inf_{\substack{\eta\in\partial f(x)\\ y\in\partial\mathcal{X}}}\{f(y) - f(x) - \langle\eta, y-x\rangle\}\end{aligned}$$

Taking the infimum over all possible values of ξ, $x_1, \ldots, x_n$ and $\theta^1, \ldots, \theta^n$ we obtain the left-hand side inequality in (iv). To obtain the remaining inequality let $w \in \mathcal{X}$ and $\xi \in \partial f(w)$. Consider $x_1, \ldots, x_n$ and $\theta^1, \ldots, \theta^n$ as before. Note that $-f(x) \leq -f(w) - \langle\xi, x-w\rangle = -f(w) - \langle\xi, x_j - w\rangle - \langle\xi, x - x_j\rangle$ for every j. We then have

$$\begin{aligned}\sum_{k=1}^{n}\theta^k f(x_k) - f(x) &\leq \sum_{k=1}^{n}\theta^k(f(x_k) - f(w) - \langle\xi, x_k - w\rangle - \langle\xi, x - x_k\rangle) \\ &\leq \sum_{k=1}^{n}\theta^k(f(x_k) - f(w) - \langle\xi, x_k - w\rangle) \\ &\leq \sup_{\substack{\xi\in\partial f(w)\\ y\in\partial\mathcal{X},\ w\in\mathcal{X}}}\{f(y) - f(w) - \langle\xi, y-w\rangle\}.\end{aligned}$$

Since this holds for every $x_1, \ldots, x_n \in \partial\mathcal{X}$ and coefficients $\theta^1, \ldots, \theta^n$ the result is now evident. $\square$

Appendix B

Results from linear algebra

Before proving Lemma 2.4.1, we need the following result.

Lemma B.0.3 *Let $j \in \{1, \ldots, d\}$, $\alpha \in \{-1, 1\}^d$ and $\rho_* > 0$. Then, the optimal value of the optimization problem*

$$\begin{array}{ll} min & \langle \alpha^j \boldsymbol{e}_j, w_2 - w_1 \rangle \\ s.t. & \left| w_2 - \frac{3\rho_*}{8\sqrt{d}}\alpha \right| \leq \frac{\rho_*}{8\sqrt{d}} \\ & |w_1| \leq \frac{\rho_*}{16\sqrt{d}} \\ & w_1, w_2 \in \mathbb{R}^d \end{array}$$

is $\frac{3}{16\sqrt{d}}\rho_*$ and it is attained at $w_1^* = \frac{\rho_*}{16\sqrt{d}}\alpha^j \boldsymbol{e}_j$ and $w_2^* = \frac{3\rho_*}{8\sqrt{d}}\alpha - \frac{\rho_*}{8\sqrt{d}}\alpha^j \boldsymbol{e}_j$.

Proof: Writing $w = (w_1; w_2)$ with $w_1, w_2 \in \mathbb{R}^d$ for any $w \in \mathbb{R}^{2d}$, consider $f, g_1, g_2 : \mathbb{R}^{2d} \to \mathbb{R}$ defined as:

$$\begin{aligned} f(w) &= \langle \alpha^j \mathbf{e}_j, w_2 - w_1 \rangle, \\ g_1(w) &= \frac{1}{2}\left(\left(\frac{\rho_*}{16\sqrt{d}}\right)^2 - |w_1|^2\right), \\ g_2(w) &= \frac{1}{2}\left(\left(\frac{\rho_*}{8\sqrt{d}}\right)^2 - \left|w_2 - \frac{3\rho_*}{8\sqrt{d}}\alpha\right|^2\right). \end{aligned}$$

Then, f, g_1, g_2 are twice continuously differentiable on $\mathbb{R}^{2d}$ and the optimization problem can be re-written as minimizing $f(w)$ over the set $\{w \in \mathbb{R}^{2d} : g_1(w) \geq 0, g_2(w) \geq 0\}$. The proof now follows by noting that the vector $w^* = (w_1^*; w_2^*) \in \mathbb{R}^{2d}$ and the Lagrange multipliers $\lambda_1^* = \frac{16\sqrt{d}}{\rho_*}$ and $\lambda_2^* = \frac{8\sqrt{d}}{\rho_*}$ are the only ones which satisfy the Karush-Kuhn-Tucker second order necessary and sufficient conditions for a strict local solution to this problem as stated in Theorem 12.5, page 343 and Theorem 12.6, page 345 in Nocedal and Wright (1999). □

B.1 Proof of Lemma 2.4.1

Without loss of generality, we may assume that $r = 1$. Let R_r be $\frac{1}{\sqrt{d}}$ and pick $\delta \in \left(0, \frac{1}{\sqrt{d}}\right)$, $\rho_* = \frac{1}{\sqrt{d}} - \delta$ and $\rho^* = \frac{2d}{1-\delta\sqrt{d}}$. Consider a matrix $Z = (z_1, \ldots, z_d) \in \mathbb{R}^{d\times d}$ with columns $z_1, \ldots, z_d \in \mathbb{R}^d$ and define the function $\tilde{\xi} : \mathbb{R}^{d\times d} \to \mathbb{R}^d$ as

$$\tilde{\xi}(Z) = \begin{vmatrix} \mathbf{e}_1 & z_2^1 - z_1^1 & \cdots & z_d^1 - z_1^1 \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{e}_d & z_2^d - z_1^d & \cdots & z_d^d - z_1^d \end{vmatrix}$$

where the bars denote the determinant and the equation is written symbolically to express that $\tilde{\xi}(Z)$ is a linear combination of the vectors $\{\mathbf{e}_j\}_{1\leq j\leq d}$ with the cofactor corresponding to the $(j, 1)$-th position as the coefficient of $\mathbf{e}_j$. This is a common notation for "generalized vector products"; see, for instance, Courant and John (1999), Section 2.4.b, page 187 for more details. Since the determinant and all cofactors can be seen as a continuous function on $\mathbb{R}^{d\times d}$, it follows that $\tilde{\xi}$ is continuous on $\mathbb{R}^{d\times d}$. Now choose $\alpha \in \{-1, 1\}^d$

and observe that

$$\begin{aligned}\tilde{\xi}(\alpha^1\mathbf{e}_1,\dots,\alpha^d\mathbf{e}_d) &= \left(\prod_{j=1}^d \alpha^j\right)\alpha,\\ \left|\tilde{\xi}(\alpha^1\mathbf{e}_1,\dots,\alpha^d\mathbf{e}_d)\right| &= \sqrt{d},\\ \langle\tilde{\xi}(\alpha^1\mathbf{e}_1,\dots,\alpha^d\mathbf{e}_d),\alpha^j\mathbf{e}_j\rangle &= \prod_{k=1}^d \alpha^k \quad \forall\ j=1,\dots,d.\end{aligned}$$

Since $\mathbb{R}^{d\times d}$ has the product topology of the d-fold topological product of $\mathbb{R}^d$ with itself, the continuity of $\tilde{\xi}$ and of $\langle\cdot,\cdot\rangle$ imply that we can find $\rho_\alpha \in \left(0, \frac{1}{\sqrt{d}} - \delta\right)$ such that if $x_j \in B(\alpha^j\mathbf{e}_j, \rho_\alpha)$ for any $j = 1,\dots,d$, $\beta = \{x_1,\dots,x_d\}$ and $X_\beta = (x_1,\dots,x_d)$, then

$$\left|\left|\tilde{\xi}(X_\beta)\right| - \sqrt{d}\right| < \delta,$$

$$\left|\frac{\tilde{\xi}(X_\beta)}{|\tilde{\xi}(X_\beta)|} - \frac{\prod_{1\le j\le d}\alpha^j}{\sqrt{d}}\alpha\right| < \delta, \tag{B.1}$$

$$\left|\left\langle\frac{\tilde{\xi}(X_\beta)}{|\tilde{\xi}(X_\beta)|}, x_j\right\rangle - \frac{\prod_{k=1}^d\alpha^k}{\sqrt{d}}\right| < \delta \quad \forall\ j=1,\dots,d. \tag{B.2}$$

Taking this into account, define

$$\xi_{\alpha,\beta} = \left(\prod_{j=1}^d \alpha^j\right)\frac{\tilde{\xi}(X_\beta)}{|\tilde{\xi}(X_\beta)|}, \text{ and } b_{\alpha,\beta} = \langle\xi_{\alpha,\beta}, x_1\rangle.$$

From the definition of the function $\tilde{\xi}$ it is straight forward to see that $\langle\xi_{\alpha,\beta}, x_j - x_1\rangle = 0\ \forall j \in \{1,\dots,d\}$, so we in fact have

$$x_1,\dots,x_d \in \mathcal{H}_{\alpha,\beta} := \{x \in \mathbb{R}^d : \langle\xi_{\alpha,\beta}, x\rangle = b_{\alpha,\beta}\}.$$

Moreover, (B.1) and (B.2) imply

$$\frac{1}{\sqrt{d}} + \delta > b_{\alpha,\beta} > \frac{1}{\sqrt{d}} - \delta > 0,$$

$$\min_{1\le j\le d}\left\{|\xi^j_{\alpha,\beta}|\right\} > \frac{1}{\sqrt{d}} - \delta > 0.$$

For simplicity, and without loss of generality (the other cases follow from symmetry), we now assume that $\alpha = \mathbf{e}$, the vector of ones. By solving the corresponding quadratic programming problems, it is not difficult to see that

$$\begin{aligned} \rho_* = \frac{1}{\sqrt{d}} - \delta < b_{\alpha,\beta} &= \inf_{\langle \xi_{\alpha,\beta}, x\rangle \geq b_{\alpha,\beta}} \{|x|\} \\ \rho^* = \frac{2d}{1-\delta\sqrt{d}} > \frac{b_{\alpha,\beta}}{\min_{1\leq j\leq d}\{|\xi^j_{\alpha,\beta}|\}} &= \sup_{\substack{\langle \xi_{\alpha,\beta}, x\rangle \leq b_{\alpha,\beta} \\ x\geq 0}} \{|x|\}. \end{aligned}$$

For the first inequality see, for instance, Exercise 16.2, page 484 of Nocedal and Wright (1999). For the second one, one must notice that $2\sqrt{d} > \frac{1}{\sqrt{d}} + \delta > b_{\alpha,\beta}$ and that the optimal value of the optimization problem must be attained at one of the vertices of the polytope $\{x \in \mathbb{R}^d_+ : \langle \xi_{\alpha,\beta}, x\rangle \leq b_{\alpha,\beta}\}$. The latter statement can be derived from the Karush-Kuhn-Tucker conditions of the problem.

The inequalities in the last display imply that $B(0,\rho_*) \subset \mathcal{H}^-_{\alpha,\beta}$ and $\{x \in \mathbb{R}^d : |x| \geq \rho^*\} \cap \mathcal{R}_\alpha \subset \mathcal{H}^+_{\alpha,\beta}$.

Finally, for $x \in B(-\alpha^j \mathbf{e}_j, \frac{1}{2}\rho_\alpha)$ we have $|x + x_j| < \rho_\alpha$ and therefore $\langle \xi_{\alpha,\beta}, x\rangle < -\langle \xi_{\alpha,\beta}, x_j\rangle + \rho_\alpha < \delta - \frac{1}{\sqrt{d}} + \rho_\alpha < 0$. We can then take any $\rho \leq \frac{1}{2}\min_{\alpha\in\{-1,1\}^d}\{\rho_\alpha\}$ to make (i)-(vi) be true. We'll now argue that by making ρ smaller, if required, (vii) also holds.

Let $B_1 = B\left(0, \frac{\rho_*}{16\sqrt{d}}\right)$, $B_2 = B\left(\frac{3\rho_*}{8\sqrt{d}}\alpha, \frac{\rho_*}{8\sqrt{d}}\right)$ and consider the functions $\varphi, \psi : \mathbb{R}^{d\times d} \to \mathbb{R}$ given by

$$\begin{aligned} \varphi(X) &= \inf_{w_1\in B_1, w_2\in B_2} \left\{ \min_{1\leq j\leq d} \left\{ (X(w_2 - w_1))^j \right\} \right\}, \\ \psi(X) &= \sup_{w_1\in B_1} \left\{ \max_{1\leq j\leq d} \left\{ (Xw_1)^j \right\} \right\}. \end{aligned}$$

Both of these functions are Lipschitz continuous with the metric induced by the $\|\cdot\|_2$-norm on $\mathbb{R}^{d\times d}$ with Lipschitz constants smaller than ρ_*. To see this, observe that

$$|X(w_2 - w_1) - Y(w_2 - w_1)| \leq \|X - Y\|_2 |w_2 - w_1| \leq \frac{9}{16}\rho_* \|X - Y\|_2$$

for all $w_1 \in B_1$, $w_2 \in B_2$ and $X, Y \in \mathbb{R}^{d\times d}$. Also, simple algebra shows that $|\min_{1\le j\le d}\{x^j\} - \min_{1\le j\le d}\{y^j\}| \le |x-y|\ \forall\, x, y \in \mathbb{R}^d$. From these assertions, one immediately gets the Lipschitz continuity of φ. Similar arguments show the same for ψ.

Let $\mathcal{I}_\alpha \in \mathbb{R}^{d\times d}$ be the diagonal matrix whose j'th diagonal element is precisely α^j. From Lemma B.0.3 it is seen that $\varphi(\mathcal{I}_\alpha) = \frac{3\rho_*}{16\sqrt{d}}$. On the other hand, it is immediately obvious that $\psi(\mathcal{I}_\alpha) = \frac{\rho_*}{16\sqrt{d}}$. Using one more time the continuity of ψ and φ and that the topology in $\mathbb{R}^{d\times d}$ is the same as the topology of the d-fold topological product of $\mathbb{R}^d$, for each $\alpha \in \{-1,1\}^d$ we can find r_α for which $X_\beta = (x_1, \ldots, x_d) \in \mathbb{R}^{d\times d}$ and $|x_j - \alpha^j \mathbf{e}_j| < r_\alpha$ for all $j = 1, \ldots, d$ imply $|\psi(X_\beta^{-1}) - \frac{\rho_*}{16\sqrt{d}}| < \frac{\rho_*}{32\sqrt{d}}$ and $|\varphi(X_\beta^{-1}) - \frac{3\rho_*}{16\sqrt{d}}| < \frac{\rho_*}{16\sqrt{d}}$. It follows that

$$
\begin{aligned}
& \inf_{\substack{t\ge 1\\ w_1\in B_1, w_2\in B_2}} \left\{ \min_{1\le j\le d} \left\{ \left(X_\beta^{-1}(w_1 + t(w_2 - w_1)) \right)^j \right\} \right\} \\
\ge\ & \inf_{\substack{t\ge 1\\ w_1\in B_1, w_2\in B_2}} \left\{ \min_{1\le j\le d} \left\{ \left(tX_\beta^{-1}(w_2 - w_1) \right)^j \right\} \right\} - \sup_{w_1\in B_1} \left\{ \max_{1\le j\le d} \left\{ \left(X_\beta^{-1} w_1 \right)^j \right\} \right\} \\
\ge\ & \varphi(X_\beta^{-1}) - \psi(X_\beta^{-1}) \ > \ \frac{\rho_*}{8\sqrt{d}} - \frac{3\rho_*}{32\sqrt{d}} \ = \ \frac{\rho_*}{32\sqrt{d}} > 0.
\end{aligned}
$$

The proof is then finished by taking $\rho \le \min_{\alpha\in\{-1,1\}^d} \left\{ r_\alpha \wedge \frac{\rho_\alpha}{2} \right\}$. □

B.2 Proof of Lemma 2.4.2

Assume again, without loss of generality, that $r = 1$. Lemma 2.4.1 (ii) and (vi) imply that $x_{\alpha^j j}, x_{-\alpha^j j} \in \{x \in \mathbb{R}^d : \langle x, \xi_\alpha\rangle \le b_\alpha\}$ for any $j = 1, \ldots, n$ and any $\alpha \in \{-1,1\}^d$. It follows that, in addition to being convex, $\cap_{\alpha\in\{-1,1\}^d}\{x \in \mathbb{R}^d : \langle \xi_\alpha, x\rangle \le b_\alpha\}$ contains $\{x_{\pm 1}, \ldots, x_{\pm d}\}$ and hence it must contain K. For the other contention, take $x \in \cap_{\alpha\in\{-1,1\}^d}\{w \in \mathbb{R}^d : \langle \xi_\alpha, w\rangle \le b_\alpha\}$ with $x \ne 0$ and any $\alpha \in \{-1,1\}^d$ for which $x \in \mathcal{R}_\alpha$. Then, $\langle \xi_\alpha, x\rangle > 0$ for otherwise we would have

$$\kappa x \in \mathcal{R}_\alpha \setminus \mathcal{H}_\alpha^+ \ \ \forall \kappa \geq 0$$

which is impossible by (v) in Lemma 2.4.1. Thus, $\mathcal{J}_x = \{\alpha \in \{-1,1\}^d : \langle \xi_\alpha, x\rangle > 0\} \neq \emptyset$ and we can define

$$r_x = \min_{\alpha \in \mathcal{J}_x} \left\{ \frac{b_\alpha}{\langle \xi_\alpha, x\rangle} \right\} \quad \text{and } \alpha_x = \operatorname*{argmin}_{\alpha \in \mathcal{J}_x} \left\{ \frac{b_\alpha}{\langle \xi_\alpha, x\rangle} \right\}.$$

Note that $r_x \geq 1$. Since β_{α_x} is a basis, there is $\theta \in \mathbb{R}^d$ such that $r_x x = \theta^1 x_{\alpha_x^1 1} + \ldots + \theta^d x_{\alpha_x^d d}$. But then,

$$b_{\alpha_x} = \langle r_x x, \xi_{\alpha_x}\rangle = \sum_{k=1}^d \theta^k \langle x_{\alpha_x^k k}, \xi_{\alpha_x}\rangle = b_{\alpha_x} \sum_{k=1}^d \theta^k$$

where the last equality follows from (ii) of Lemma 2.4.1 and therefore $\theta^1 + \ldots + \theta^d = 1$. Now assume that $\theta^j < 0$ for some $j \in \{1, \ldots, d\}$ and set $\gamma_x \in \{-1,1\}^d$ with $\gamma_x^k = \alpha_x^k$ for $k \neq j$ and $\gamma_x^j = -\alpha_x^j$. But then, $\sum_{k \neq j} \theta^k = 1 - \theta^j > 1$, $\langle x_{\alpha_x^k k}, \xi_{\gamma_x}\rangle = b_{\gamma_x}$ for $k \neq j$ and $\langle x_{\alpha_j^j j}, \xi_{\gamma_x}\rangle < 0$ by (ii) and (vi) in Lemma 2.4.1. Therefore,

$$\begin{aligned} \langle r_x x, \xi_{\gamma_x}\rangle &= \theta^j \langle x_{-\alpha_x^j j}, \xi_{\gamma_x}\rangle + \sum_{k \neq j} \theta^k \langle x_{\alpha_x^k k}, \xi_{\gamma_x}\rangle && \text{(B.3)} \\ &> \sum_{k \neq j} \theta^k \langle x_{\alpha_x^k k}, \xi_{\gamma_x}\rangle > b_{\gamma_x} && \text{(B.4)} \end{aligned}$$

which is impossible because it contradicts the definition of r_x. Hence, $\theta \geq 0$ and we have $r_x x \in Conv\,(\beta_{\alpha_x})$. Note that since 0 belongs in the interior of $\cap_{\alpha \in \{-1,1\}^d} \{w \in \mathbb{R}^d : \langle \xi_\alpha, w\rangle \leq b_\alpha\}$, there there is $\kappa > 0$ such that $-\kappa x \in \cap_{\alpha \in \{-1,1\}^d} \{w \in \mathbb{R}^d : \langle \xi_\alpha, w\rangle \leq b_\alpha\}$. Applying the same arguments as before to $-\kappa x$ instead of x, we can find $\tilde{r}_x > 0$ and $\tilde{\alpha}_x \in \{-1,1\}^d$ such that $-\tilde{r}_x x \in Conv\,(\beta_{\tilde{\alpha}_x})$. It follows that $-\tilde{r}_x x, r_x x \in K$ and therefore $0, x \in K$ since $r_x \geq 1$. Hence, we have proved (i).

To prove (ii), note that $A := \cap_{\alpha\in\{-1,1\}^d}\{w \in \mathbb{R}^d : \langle \xi_\alpha, w\rangle < b_\alpha\}$ is open and, by (i), it is contained in K. Thus, $A \subset K^\circ$. That $K^\circ \subset A$ follows from the fact that if $x \in K \setminus A$, then $\langle \xi_\alpha, x\rangle = b_\alpha$ for some $\alpha \in \{-1,1\}^d$, which implies that $B(x,\tau) \cap \mathrm{Ext}(K) \neq \emptyset$ for all $\tau > 0$ and hence $x \notin K^\circ$.

It is then obvious that (iv) follows from the identity $\partial K = \overline{K} \setminus K^\circ$ and the fact that K is closed.

Pick any $\alpha \in \{-1,1\}^d$ and observe that (ii) and (vi) from Lemma 2.4.1 imply that for any $\gamma \in \{-1,1\}^d$ we have

$$\langle \xi_\gamma, x_{\alpha^k k}\rangle \begin{cases} = b_\gamma & \text{if } \gamma^k = \alpha^k \\ < 0 \le b_\gamma & \text{if } \gamma^k = -\alpha^k \end{cases}$$

which by (iv) of this lemma show that

$$x_{\alpha^j j} \in \{w \in \mathbb{R}^d : \langle \xi_\alpha, w\rangle = b_\alpha\} \cap \left(\cap_{\gamma\in\{-1,1\}^d}\{w \in \mathbb{R}^d : \langle \xi_\gamma, w\rangle \le b_\gamma\}\right)$$

for all $\alpha \in \{-1,1\}^d$ and $j = 1,\ldots,d$. Since the sets on the right-hand side of the last display are all convex we can conclude that

$$Conv\left(x_{\alpha^1 1},\ldots,x_{\alpha^j j}\right) \subset \{w \in \mathbb{R}^d : \langle \xi_\alpha, w\rangle = b_\alpha\} \cap \Big(\cap_{\gamma\in\{-1,1\}^d}\{w \in \mathbb{R}^d : \langle \xi_\gamma, w\rangle \le b_\gamma\}\Big)$$

for all $\alpha \in \{-1,1\}^d$. Thus, $\bigcup_{\alpha\in\{-1,1\}^d} Conv\left(x_{\alpha^1 1},\ldots,x_{\alpha^j j}\right) \subset \partial K$. Finally, take $x \in \partial K$. Then, there is $\alpha_x \in \{-1,1\}^d$ such that $\langle \xi_{\alpha_x}, x\rangle = b_{\alpha_x}$. Since β_{α_x} is a basis we can again find $\theta \in \mathbb{R}^d$ such that $x = \theta^1 x_{\alpha_x^1 1} + \ldots + \theta^d x_{\alpha_x^d d}$. Just as before, $\langle \xi_{\alpha_x}, x_{\alpha_x^j j}\rangle = b_{\alpha_x}$ implies that $\sum \theta^j = 1$. And again, if $\theta^j < 0$ for some j, we can take $\gamma_x \in \{-1,1\}^d$ with $\gamma_x^k = \alpha_x^k$ for $k \neq j$ and $\gamma_x^j = -\alpha_x^j$ and arrive at a contradiction with similar arguments to those used in (B.3) and (B.4). This shows that $x \in Conv\left(\beta_{\alpha_x}\right)$ and completes the proof as (v) and (vi) are direct consequences of $(i) - (iv)$ and Lemma 2.4.1. $\square$

B.3 Proof of Lemma 2.4.3

Let $r \in (0, \frac{1}{d-2})$ if $d \geq 3$ and $r > 0$ if $d \leq 2$. Since the geometric properties of any rectangle depend only on the direction and magnitude of the diagonal, we may assume without loss of generality that $b > 0$ and that $a = \frac{r}{1+r}b$. This is because we can define $\tilde{b} = (1+r)(b-a) > 0$ and $\tilde{a} = a - r(b-a)$ to obtain $[a, b] = \tilde{a} + \left[\frac{r}{r+1}\tilde{b}, \tilde{b}\right]$. For any $\alpha \in \{-1, 1\}^d$, define $\alpha_j = \alpha - 2\alpha^j \mathbf{e}_j \in \mathbb{R}^d$ and $w_\alpha = z_\alpha + r(z_\alpha - z_{-\alpha})$. Additionally, define the functions $\psi_\alpha, \varphi_\alpha : \mathbb{R}^{d\times d} \times \mathbb{R}^d \to \mathbb{R}$ by

$$
\begin{aligned}
\psi_\alpha(\Theta, \theta) &= \langle \mathbf{e}, \Theta(z_\alpha - \theta) \rangle \\
\varphi_\alpha(\Theta, \theta) &= \min_{1 \leq j \leq d} \left\{ (\Theta(z_\alpha - \theta))^j \right\}.
\end{aligned}
$$

Considering $\mathbb{R}^{d\times d}$ with the topology generated be the $\|\cdot\|_2$ norm and $\mathbb{R}^{d\times d} \times \mathbb{R}^d$ with the product topology, it is easily seen that both functions defined in the last display are continuous. Now, let $W_\alpha \in \mathbb{R}^{d\times d}$ be the matrix whose j'th column is precisely $w_{\alpha_j} - w_\alpha$. It is not difficult to see that $\psi_\alpha(W_\alpha^{-1}, w_\alpha) = \frac{dr}{1+2r} < 1$ and $\varphi_\alpha(W_\alpha^{-1}, w_\alpha) = \frac{r}{1+2r} > 0$. For instance, one can check that for $\alpha = -\mathbf{e}$, one has $w_\alpha = 0$ and $w_{\alpha_j} = \frac{1+2r}{1+r} b^j \mathbf{e}_j$ and the result is now evident. By symmetry, the same is true for any $\alpha \in \{-1, 1\}^d$. Therefore, for any $\alpha \in \{-1, 1\}^d$ there is ρ_α such that whenever $|x_{\alpha_j} - w_{\alpha_j}| < \rho_\alpha \ \forall\ j = 1, \ldots, d$ and X_α is the matrix whose j'th column is $x_{\alpha_j} - x_\alpha$, we get

$$
\psi_\alpha(X_\alpha^{-1}, x_\alpha) < 1, \tag{B.5}
$$

$$
\varphi_\alpha(X_\alpha^{-1}, x_\alpha) > 0. \tag{B.6}
$$

Letting $\rho = \min_{\alpha \in \{-1,1\}^d} \{\rho_\alpha\}$ completes the proof as (B.5) and (B.6) imply $z_\alpha \in Conv\,(x_\alpha, x_{\alpha_1}, \ldots, x_{\alpha_d})^\circ$. $\square$