

Density estimation under qualitative assumptions in higher dimensions

by Wolfgang Polonik

Universität Heidelberg

Abstract:

We study a method for estimating a density f in \mathbf{R}^d under assumptions which are of qualitative nature. The resulting density estimator can be considered as a generalization of the Grenander estimator for monotone densities. The assumptions on f are given in terms of the density contour clusters $\Gamma(\lambda) = \{x : f(x) \geq \lambda\}$. We assume that for all $\lambda \geq 0$ the sets $\Gamma(\lambda)$ lie in a given class \mathbb{C} of measurable subsets of \mathbf{R}^d . By choosing \mathbb{C} appropriately it is possible to model for example monotonicity, symmetry or multimodality. The main mathematical tool for proving consistency and rates of convergence of the density estimator is empirical process theory. It will turn out that the rates depend on the richness of \mathbb{C} measured by metric entropy.

1. Introduction

We provide a method for estimating a density f in \mathbf{R}^d under qualitative assumptions like monotonicity, symmetry, modality. More generally, we assume that the density contour clusters $\Gamma(\lambda) = \Gamma_f(\lambda) = \{x : f(x) \geq \lambda\}$, $\lambda > 0$, all lie in a given class \mathbb{C} of measurable subsets of \mathbf{R}^d . For example, if we choose $\mathbb{C} = \mathbb{C}_0 = \{ [0,x], x \geq 0 \}$ then the class of densities with $\Gamma(\lambda) \in \mathbb{C}_0$ is the class of all nonincreasing leftcontinuous densities on the positive real line. Hence, the choice of \mathbb{C} can be interpreted as choosing a statistical model.

The density estimator studied in this paper is based on estimators of the sets $\Gamma(\lambda)$, so that density estimation as considered here can be viewed as a certain two-step procedure: first estimate the density contour clusters and then estimate the density by means of the estimated density contour clusters.

For the moment consider the special case of estimating a monotone density on the real line. If the mode is fixed, then the maximum likelihood estimator of a , lets say, decreasing left continuous density is well known. It is the so-called Grenander estimator which is defined as the slope (more precisely, the left-hand derivative) of the smallest concave majorant of the empirical distribution function. Hence, the Grenander estimator is obtained by first estimating the distribution function under the assumption that it is concave. This estimator is the smallest concave majorant of the empirical distribution function. Then the density is estimated by the slope of the concave majorant. It turns out, that for the special choice of $\mathbb{C} = \mathbb{C}_0$ our density estimator coincides with the Grenander estimator. The connection of estimating density contour clusters in \mathbb{C}_0 and the concave majorant of the empirical distribution function will be given below.

Another density estimator known in the literature which is also constructed by estimating density contour clusters has been considered by Sager [12]. In our notation he assumed that all sets $\Gamma(\lambda)$, $\lambda > 0$, lie in \mathbb{E}^d , the class of closed convex sets in \mathbf{R}^d . As estimators for $\Gamma(\lambda)$ he used minimal volume sets in \mathbb{E}^d . A minimum volume set in \mathbb{E}^d to the parameter α is defined to be the smallest set in \mathbb{E}^d which contains at least empirical mass α . Sager constructed a density estimator out of a nested sequence $\{C_n\}$ of minimal volume sets by putting slices $C_n \times [a_n, b_n]$ one on the top of the other. The difficulty there is to choose the thickness of the slices, $b_n - a_n$, in an appropriate way. However, by construction the estimators $\{C_n\}$ of the density contour clusters are nested. Note, that in general this is not the case for minimal volume sets to different parameters α . In general, minimal volume sets may overlap. They even may not intersect.

The density estimators considered in the present paper can also be visualized as putting slices constructed out of minimum volume sets one on top of the other. But in contrast to the estimator of Sager the thickness of the slices as well as the parameters α corresponding to the minimum volume sets depend on the data and come out of the procedure automatically. However, the problem that the estimators of the density contour clusters may overlap also appears in our context. This will be discussed below.

Before we give the organization of the paper we shortly discuss the assumption “ $\Gamma(\lambda) \in \mathbb{C} \forall \lambda > 0$ ”. An equivalent formulation for this assumption is “ $f \in F_{\mathbb{C}}$ ”, where

$$F_{\mathbb{C}} = \{ f : \mathbf{R}^d \rightarrow [0, \infty), \int f(x) dx = 1, \Gamma_f(\lambda) \in \mathbb{C} \forall \lambda > 0 \}.$$

First consider the case $d = 1$. As already mentioned, for $\mathbb{C} = \mathbb{C}_0 = \{ [0, x], x \geq 0 \}$, $F_{\mathbb{C}}$ equals the class of all monotone decreasing (left-continuous) densities on the real line starting at zero. Let \mathfrak{I}_k denote the class which consists of all unions of at most $k \geq 1$ closed intervals on the real line. $F_{\mathfrak{I}_1}$ defines the class of unimodal densities on the real line. The class $F_{\mathfrak{I}_2} \setminus F_{\mathfrak{I}_1}$ consists of densities with at most two modes, more general $F_{\mathfrak{I}_k} \setminus F_{\mathfrak{I}_{k-1}}$ consists of densities with at most k modes. In order to model modality in higher dimensions there is no such natural choice as the class of intervals in the one-dimensional case. In principle every class which consists of connected sets can be used to model unimodality. Standard examples are given by the classes of all closed balls, ellipsoids and convex sets in \mathbf{R}^d , denoted by \mathbf{B}^d , \mathbf{E}^d and \mathbf{C}^d , respectively. In the multimodal case the density contours become more complicated. Here classes which can be constructed out of the convex sets by means of finitely many set-theoretic operations \cap , \cup , c seem to be appropriate (see Polonik [10] for a discussion). Taking \mathbb{C} as the class of all balls with midpoint zero leads to the class of spherically symmetric densities on \mathbf{R}^d with center zero.

The paper is organized as follows: In Section 2 we define the estimators of the density contour clusters and the density estimator itself. Furthermore we show the above mentioned connection to the Grenander estimator. Section 3 contains asymptotic results about the density estimator. By means of empirical process theory we show consistency and give rates of convergence. It will turn out, that \mathbb{C} enters the rates through its richness measured by metric entropy. The well known rate of the Grenander estimator ($n^{-1/3}$ for a density with no flat part and $n^{-1/2}$ for an underlying uniform

distribution) will be re-derived up to a log-term. They come up only through the fact that the corresponding class \mathbb{C}_0 is a so called VC-class. After some concluding remarks (Section 4), the proofs of the results of Sections 2 and 3 are given in Section 5.

2. The density estimator

For any set C let $\mathbb{1}_C$ denote the corresponding indicator function. The following equality holds for any density f :

$$(2.1) \quad f(x) = \int \mathbb{1}_{\Gamma(\lambda)}(x) d\lambda \quad \forall x \in \mathbf{R}.$$

For many class \mathbb{C} there exists an estimator of $\Gamma(\lambda)$, denoted by $\Gamma_{n,\mathbb{C}}(\lambda)$, which lies in \mathbb{C} (see Polonik [10, 11]). This estimator is called *empirical generalized λ -cluster (in \mathbb{C})*. (The definition of $\Gamma_{n,\mathbb{C}}(\lambda)$ will be given below). We define the plug in estimator of f as

$$(2.2) \quad f_{n,\mathbb{C}}(x) = \int \mathbb{1}_{\Gamma_{n,\mathbb{C}}(\lambda)}(x) d\lambda \quad \forall x \in \mathbf{R}.$$

For $d = 1$ and $\mathbb{C} = \mathbf{I}_k$, $k \in \mathbf{N}$, this estimator has implicitly been used by Müller and Sawitzki [7] in order to determine bootstrap critical values for tests of multimodality. They draw bootstrap samples out of distributions determined by f_{n,\mathbf{I}_k} .

The empirical generalized λ -clusters in \mathbb{C} :

Let F be a distribution on \mathbf{R}^d with Lebesgue density f . It is easy to see that $\Gamma(\lambda)$ maximizes the signed measure $H_\lambda = F - \lambda \text{Leb}$ over all measurable sets, i.e.

$$H_\lambda(\Gamma(\lambda)) = \sup \{ H_\lambda(C), C \text{ measurable} \}.$$

As a function of λ the maximal value $E(\lambda) = H_\lambda(\Gamma(\lambda))$ is called *excess mass functional*. It has

been introduced by Müller and Sawitzki [7] and can be used for investigating the modality of a distribution (see Müller and Sawitzki[7, 8], Hartigan [4], Nolan [9], Polonik [10]).

Let (Ω, A, P) denote the underlying probability space and let X_1, X_2, \dots be i.i.d. random vectors in \mathbf{R}^d with distribution F . Furthermore, let F_n denote the empirical distribution function of the first n observations. Define the empirical version of H_λ as

$$H_{n,\lambda} = F_n - \lambda \text{ Leb.}$$

The supremum of $H_{n,\lambda}$ over all measurable sets equals one and is attained at $\{X_1, X_2, \dots, X_n\}$. Hence, in order to obtain a reasonable estimator of $\Gamma(\lambda)$ by means of $H_{n,\lambda}$ one has to restrict the maximization to certain subclasses \mathbb{C} of subsets of \mathbf{R}^d .

Definition: Let \mathbb{C} be a class of measurable subsets of \mathbf{R}^d . Any set $\Gamma_{n,\mathbb{C}}(\lambda) \in \mathbb{C}$, such that

$$H_{n,\lambda}(\Gamma_{n,\mathbb{C}}(\lambda)) = \sup_{C \in \mathbb{C}} H_{n,\lambda}(C)$$

is called an empirical generalized λ -cluster in \mathbb{C} .

We called those sets *generalized* because they need not be connected as one perhaps would expect for “clusters”.

In the one-dimensional case, more precisely, for $\mathbb{C} = \mathfrak{I}_k$, Müller and Sawitzki [8] gave consistency results for the empirical generalized λ -clusters. In a more parametric setting Nolan [9] studied the sets $\Gamma_{n,\mathbb{C}}(\lambda)$ for $\mathbb{C} = \mathfrak{E}^d$. She gave consistency results and asymptotic distributions for the corresponding finite-dimensional parameters. Under more general conditions on \mathbb{C} (analogous to those used in the present paper; see results in Section 3) the empirical generalized λ -clusters $\Gamma_{n,\mathbb{C}}(\lambda)$ have been studied in Polonik [10, 11].

Remark 2.1: (i) Note that the empirical generalized λ -clusters $\Gamma_{n,\mathbb{C}}(\lambda)$ are *minimum volume sets* in \mathbb{C} : by definition they have the smallest volume among all other sets in \mathbb{C} which carry the same empirical mass $F_n(\Gamma_{n,\mathbb{C}}(\lambda))$. However, this mass itself is also random.

(ii) Of course it is necessary for $\Gamma_{n,\mathbb{C}}(\lambda)$ to be a d_1 -consistent estimator of $\Gamma(\lambda)$ that the maximizing value of H_λ is unique up to F -nullsets. This is the case if and only if f has no flat part

at the level λ , i.e. if $\text{Leb}\{x: f(x) = \lambda\} = 0$. Since we shall use such results about the estimation of $\Gamma(\lambda)$ assumptions about flat parts of f will enter the theorems about rates of convergence of the density estimator given in Section 3.

The empirical generalized λ -clusters exist for standard classes \mathbb{C} like \mathbf{B}^d , \mathbf{E}^d or \mathbf{C}^d , which denote the classes of all closed balls, ellipsoids and convex sets, respectively, in \mathbf{R}^d . For the case of convex sets this can be seen easily, because one only has to consider those convex polygons with vertices in the observations. And these are (for fixed n) only finitely many. For $\mathbb{C} = \mathbf{B}^d$ or \mathbf{E}^d the maximization of $H_{n,\lambda}$ can also be reduced to a finite class of sets.

We shall assume in all of that what follows that:

(A) *For all $\lambda \geq 0$ there exist an empirical generalized λ -cluster in \mathbb{C} and $\emptyset \in \mathbb{C}$.*

The sets $\Gamma_{n,\mathbb{C}}(\lambda)$ need not be uniquely determined. Note that it even may happen that for a fixed level λ there exist empirical generalized λ -clusters which carry different empirical mass and therefore also have different Lebesgue measure. However, the following properties hold:

Proposition 2.2:

(a) *For each $\lambda \geq 0$ choose a set $\Gamma_{n,\mathbb{C}}(\lambda)$. For each such choice the function $\lambda \rightarrow \text{Leb}(\Gamma_{n,\mathbb{C}}(\lambda))$, $\lambda \geq 0$ is monotonically decreasing and piecewise constant with at most $n + 1$ jumps. Moreover, the sets $\Gamma_{n,\mathbb{C}}(\lambda)$ can be chosen such that $\lambda \rightarrow \Gamma_{n,\mathbb{C}}(\lambda)$ is piecewise constant.*

(b) *There exists a level $\lambda_{n,\max} \geq 0$ such that $\text{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)) = 0$ for $\lambda > \lambda_{n,\max}$*

In order to obtain a reasonable representation for our density estimator we suppose the sets $\Gamma_{n,\mathbb{C}}(\lambda)$ to be chosen such that:

(α) the function $\lambda \rightarrow \Gamma_{n,\mathbb{C}}(\lambda)$, $\lambda \geq 0$, is piecewise constant. Denote the (random) levels where this function has jumps by $0 = \lambda_0 < \lambda_1 < \dots < \lambda_{k_n} = \lambda_{n,\max}$, $k_n \leq n$.

(β) For every fixed $\mu \geq 0$ the Lebesgue measure of $\Gamma_{n,\mathbb{C}}(\mu)$ is minimal among all empirical generalized μ -clusters.

Furthermore, we define with $\lambda_{n,\max}$ of Proposition 2.2 (b):

(γ) $\Gamma_{n,\mathbb{C}}(\lambda) = \emptyset$ for $\lambda > \lambda_{n,\max}$.

Because of assumption (**A**) there always exist choices of $\Gamma_{n,\mathbb{C}}(\lambda)$, $\lambda > 0$, such that (α) and (β) are satisfied. This follows from the proof of Proposition 2.2 which is given in Section 5. Note that (α) and (β) do not affect the results given below, because these results hold for any choice of sets $\Gamma_{n,\mathbb{C}}(\lambda)$.

Note that in general the empty set is not an empirical generalized λ -cluster in \mathbb{C} for $\lambda > \lambda_{n,\max}$. However, for all standard classes \mathbb{C} mentioned in this paper (γ) does also not affect the asymptotic results given in Section 3 (see Section 4 for more thorough discussion on (γ)).

(α), (β) and (γ) are supposed to hold in all of that what follows. The next proposition shows that in many situations $f_{n,\mathbb{C}}$ automatically is a probability density.

Proposition 2.3: *If $\lambda_{n,\max} > 0$, then we have*

$$\int f_{n,\mathbb{C}}(x) dx = F_n(\Gamma_{n,\mathbb{C}}(0)).$$

Remark: The value of $\lambda_{n,\max}$ (especially if it is > 0) depends on the class \mathbb{C} , the sample size n and on the underlying distribution. For example, if $\mathbb{C} = \mathbb{C}^2$ and $n \leq 3$ then $\lambda_{n,\max} = 0$, because for $\lambda = 0$ the convex hull of the sample is a empirical generalized λ -cluster, and for $\lambda > 0$ we either have a datapoint itself or a line connecting two datapoints as empirical λ -clusters. Both have Lebesgue measure zero, and hence, by definition of $\lambda_{n,\max}$ we have $\lambda_{n,\max} = 0$. For $n \geq 4$ this does not happen if not three or more points lie on a line. And this in turn happens with probability zero if the underlying distribution is continuous.

Because of Proposition 2.2 we call a class \mathbb{C} *normalizing* if $F_n(\Gamma_{n,\mathbb{C}}(0)) = 1$. Examples for normalizing classes are $\mathbb{C} = \mathbb{C}_0$, $\mathbf{1}_k$ for $d = 1$, and $\mathbb{C} = \mathbf{B}^d, \mathbf{E}^d, \mathbf{C}^d$ for higher dimensions. Also

those classes which can be constructed out of the former classes by finitely many set theoretic operations \cap , \cup , c are normalizing.

Under (α) , (β) , (γ) the density estimator $f_{n,\mathbb{C}}(x)$ can be written as

$$(2.3) \quad f_{n,\mathbb{C}}(x) = \sum_{j=0}^{k_n} (\lambda_{j+1} - \lambda_j) \mathbb{I}_{\Gamma_{n,\mathbb{C}}(\lambda_j)}(x).$$

Hence, if in addition the sets $\Gamma_{n,\mathbb{C}}(\lambda_j)$, $j = 0, \dots, k_n$ are monotonically decreasing for inclusion, i.e. $\Gamma_{n,\mathbb{C}}(\lambda_{j+1}) \subset \Gamma_{n,\mathbb{C}}(\lambda_j)$, then $f_{n,\mathbb{C}}$ can be visualized as putting the slices $\Gamma_{n,\mathbb{C}}(\lambda_j) \times [\lambda_j, \lambda_{j+1}]$ one on top of the other.

However, unfortunately the monotonicity of the empirical generalized λ -clusters need not hold. In this case the density contour clusters of $f_{n,\mathbb{C}}$ need not necessarily lie in \mathbb{C} , so that $f_{n,\mathbb{C}}$ *does in general not lie in the model class $F_{\mathbb{C}}$* . But if the model is correct, i.e. if $f \in F_{\mathbb{C}}$, then as we shall show in Section 3, $f_{n,\mathbb{C}}$ converges to f as n tends to infinity, so that at least asymptotically $f_{n,\mathbb{C}}$ does lie in $F_{\mathbb{C}}$. However, there exist situations where the monotonicity of the empirical generalized λ -clusters holds automatically. For example, consider $\mathbb{C} = \mathbb{C}_0 = \{ [0,x], x \geq 0 \}$. In this class the empirical generalized λ -clusters are monotonically decreasing for inclusion, because they all start in zero and their Lebesgue measures are decreasing (Proposition 2.2 (a)).

$f_{n,\mathbb{C}}$ and the Grenander estimator

As already mentioned, $F_{\mathbb{C}_0}$ is the class of all monotone decreasing, left-continuous densities on the real line, starting at zero. In this model there exists a well-known estimator for the density: the Grenander estimator \hat{f}_n (Grenander [5]), which is the maximum likelihood estimator of f in $F_{\mathbb{C}_0}$. It has been shown by Grenander that \hat{f}_n is given by the left-continuous of the smallest concave majorant of F_n , denoted by F_n^* . Surprisingly, in this special situation f_{n,\mathbb{C}_0} and the Grenander estimator \hat{f}_n coincide. This can be seen as follows. Let

$$U_n(\lambda) := \inf\{ t \geq 0 : F_n(t) - \lambda t \text{ is maximal } \}.$$

Here $F_n(\cdot)$ denotes the empirical distribution function of n i.i.d.-observations drawn from F . We

use the symbol “ F_n ” for the empirical distribution function and for the empirical distribution itself. U_n has the property that

$$\hat{f}_n(x) \leq \lambda \iff U_n(\lambda) \leq x.$$

(This has already been used in Groeneboom [6]). Furthermore we have $\Gamma_{n, \mathbb{C}_0}(\lambda) = [0, t_\lambda]$ where $t_\lambda = \operatorname{argmax}_{t \geq 0} \{F_n([0, t]) - \lambda \operatorname{Leb}([0, t])\} = \operatorname{argmax}_{t \geq 0} \{F_n(t) - \lambda t\}$. Together with (β) it follows that $t_\lambda = U_n(\lambda)$. Hence, by using the monotonicity of the empirical generalized λ -clusters we obtain that

$$f_{n, \mathbb{C}}(x) \leq \lambda \iff x \notin \Gamma_{n, \mathbb{C}}(\lambda) \iff t_\lambda \leq x \iff U_n(\lambda) \leq x,$$

and hence f_{n, \mathbb{C}_0} and \hat{f}_n coincide.

Note that $f_{n, \mathbb{C}}$ has jumps at $x = t_{\lambda_j}$, $j = 0, \dots, k_n$, where the λ_j are defined in (α) above. By definition of $\Gamma_{n, \mathbb{C}}(\lambda)$ we have $t_{\lambda_j} = X_k$ for some $k = k(j)$. On the other hand, the Grenander estimator has jumps at those points where the slope of F_n^* changes, or in other words, where F_n and F_n^* coincide. Hence, the sets $\{\Gamma_{n, \mathbb{C}_0}(\lambda_j), j = 0, \dots, k_n\}$ can be constructed by first deriving F_n^* and then choosing those points where the slope of F_n^* changes as endpoints of the interval starting at zero. The corresponding levels λ_j are given by the (left-sided) derivative of F_n^* in t_{λ_j} .

3. Asymptotic results

In this section consistency results and rates of convergence for $f_{n, \mathbb{C}}$ will be given in terms of the L_1 -distance. Let $d_1(f, g)$ denote the L_1 -distance of two functions, i.e. $d_1(f, g) = \int |f(x) - g(x)| dx$. The L_1 -distance of two sets C, D is defined as the L_1 -distance of the corresponding indicator functions, so that $d_1(C, D) = \operatorname{Leb}(C \Delta D)$, where “ Δ ” denotes the symmetric difference and “ Leb ” the Lebesgue measure.

In order to avoid measurability considerations we define for any function $f : \Omega \rightarrow \mathbf{R}$ the *measurable cover function* f^* as the smallest measurable function from Ω to \mathbf{R} lying everywhere above f . Of course, if f is measurable, then $f^* = f$. Furthermore, let P^* denote *outer probability*. Note that for any $\alpha > 0$ we have

$$P^*(f > \alpha) = P(f^* > \alpha).$$

See for example Dudley [2] for more details. We need the following definition:

Definition: \mathcal{C} is called a *Glivenco-Cantelli (GC)-class* for F , or a *GC(F)-class*, if with probability 1

$$\sup_{C \in \mathcal{C}} |F_n(C) - F(C)|^* \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

The classes $\mathcal{C} = \mathcal{C}_0, \mathbf{1}_k, \mathbf{B}^d, \mathbf{E}^d$ are GC(F)-classes for all F . The class \mathcal{E}^d is a GC(F)-class if for example F has a bounded Lebesgue density (see Eddy and Hartigan [3] for a characterization of the GC(F)-property of \mathcal{E}^d). Moreover, all classes which can be constructed out of GC-classes by means of a finite number of the set theoretic operations $\cap, \cup, ^c$ are GC-classes. As in Alexander [1] we call such classes *k-constructible*: more precisely, a class \mathcal{C} in a measurable space (X, \mathcal{A}) is called *k-constructible* from a GC-class \mathcal{D} , if there exists a function φ from \mathbb{D}^k to \mathcal{A} constructed from $\cap, \cup, ^c$ such that $\mathcal{C} \subset \varphi(\mathbb{D}^k)$. For example, the class of all at most k -sided polygons in \mathbf{R}^2 is k -constructible from the class of all halfplanes, since they can be written as an intersection of at most k halfplanes.

Theorem 3.1 (Consistency): *Suppose that \mathcal{C} is normalizing. If $f \in F_{\mathcal{C}}$, then there exists a real-valued (non-random) function $A = A(\eta, L)$, depending on f , with $A(\eta, L) \rightarrow 0$ as $\eta \rightarrow 0$ and $L \rightarrow \infty$, such that for all $L, \eta > 0$ with $L \geq \eta$ we have*

$$(3.1) \quad d_1(f_{n, \mathcal{C}}, f) \leq 2 \eta^{-1} \int_{\eta}^L [(F_n - F)(\Gamma_{n, \mathcal{C}}(\lambda)) - (F_n - F)(\Gamma(\lambda))] d\lambda + A(\eta, L).$$

Hence, if in addition \mathcal{C} is a GC(F)-class then we have with probability 1 that

$$d_1(f_{n, \mathcal{C}}, f)^* \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Remark: It can be seen from the proof of (3.1) that $A(\eta, L) \leq 4 \int_0^{2\eta} \varphi(\lambda) d\lambda + 2 \int_L^{\infty} \varphi(\lambda) d\lambda$, where $\varphi(\lambda) := \text{Leb}\{x : f(x) \geq \lambda\}$. If f is bounded by M , say, then the last integral equals zero for L

$> M$ (because $\varphi(\lambda) = 0$ for $\lambda > M$). Hence, if we denote $A(\eta) = A(\eta, L)$, $L > M$, then the behaviour of $A(\eta)$ as $\eta \rightarrow 0$ reflects the tail behaviour of f . It will be shown below that this behaviour also is crucial for the rate of convergence of $f_{n, \mathbb{C}}$.

Moreover, it can be seen from (3.1) that the convergence is uniform over a certain class of densities if $A(\eta, L)$ tends to zero uniformly over this class.

Rates of convergence:

We give rates of convergence for two different types of classes \mathbb{C} . The first type are so-called Vapnik-Cervonenkis (VC)-classes, and the second type are classes which satisfy a certain entropy condition (see (3.3) below).

Note that the L_1 -distance of f_n and f can be bound in terms of the L_1 -distance of the sets $\Gamma_{n, \mathbb{C}}(\lambda)$ and $\Gamma(\lambda)$. We have by means of Fubini's theorem:

$$\begin{aligned}
 d_1(f_{n, \mathbb{C}}, f) &= \int \left| \int \mathbb{1}_{\Gamma_{n, \mathbb{C}}(\lambda)}(x) - \mathbb{1}_{\Gamma(\lambda)}(x) \, d\lambda \right| dx \leq \int \text{Leb}(\Gamma_{n, \mathbb{C}}(\lambda) \Delta \Gamma(\lambda)) \, d\lambda \\
 (3.2) \qquad \qquad \qquad &= \int d_1(\Gamma_{n, \mathbb{C}}(\lambda), \Gamma(\lambda)) \, d\lambda.
 \end{aligned}$$

Note that this inequality is an equality if the empirical generalized λ -clusters are nested so that they are density contour clusters of $f_{n, \mathbb{C}}$.

(3.2) shows, that results about the behaviour of $d_1(\Gamma_{n, \mathbb{C}}(\lambda), \Gamma(\lambda))$ can be used to obtain results about $d_1(f_n, f)$. Now, estimation of $\Gamma(\lambda)$ by $\Gamma_{n, \mathbb{C}}(\lambda)$ is “critical” if f has flat parts at level λ (see Remark 2.1). Because of (3.2) such levels might also be critical for the estimation of the density itself. From this point of view we define “critical levels” as follows: Let $\varphi(\lambda) = \varphi_f(\lambda) = \text{Leb}(\Gamma(\lambda)) = \text{Leb}\{x: f(x) \geq \lambda\}$.

Definition: A level $\lambda > 0$ is called a critical level (of f), if φ is not differentiable at λ .

Note that if f has a flat part at a level λ , i.e. $\text{Leb}\{x: f(x) = \lambda\} > 0$, then φ is not continuous at λ and hence such levels are critical.

Now we consider the case that \mathbb{C} is a VC-class. VC-classes are defined through a combinatorical

property as follows: Let D be a finite set in \mathbf{R}^d . A class \mathbb{C} is said to “shatter” D iff every $B \subset D$ is of the form $C \cap D$ for some $C \in \mathbb{C}$. If there exists a number $k \in \mathbf{N}$, such that \mathbb{C} shatters no set which consists of k elements, then \mathbb{C} is called a VC-class and the minimal k with that property is called the *index* of \mathbb{C} . Examples for VC-classes are the classes of intervals, \mathfrak{I}_1 , and more general, the classes \mathfrak{I}_k , which consist of all unions of at most k intervals. Every subclass of a VC-class also is a VC-class as for example the class \mathbb{C}_0 (as a subclass of \mathfrak{I}_1) which corresponds to the Grenander estimator (see above). Examples for VC-classes in higher dimensions are the classes of all halfplanes and the classes \mathfrak{B}^d and \mathfrak{E}^d .

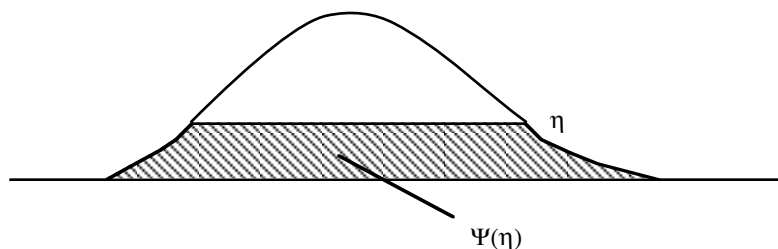
For the proofs of the theorems given below we shall use results of Alexander [1] about the behaviour of the set- and function-indexed empirical process. For that reason we shall also use some of his terminology:

Alexander considers classes of functions or sets, respectively, which satisfy a certain measurability condition which he called “*n-deviation measurable*”. Here we shall not give this definition and the underlying construction of the empirical measure, because all the standard VC-classes which we are interested in (the classes of balls, ellipsoids in \mathbf{R}^d and classes which are k -constructible out of the former classes as defined above) satisfy this measurability condition. Furthermore, we call \mathbb{C} a *(v,k)-constructible VC-class*, if \mathbb{C} is k -constructible from a VC-class \mathbb{D} whose index is smaller than or equal to v .

The tail behaviour of the underlying density f is crucial for the rates of convergence of the density estimator. This tail behaviour will here be measured in terms of the function

$$\Psi(\eta) = \Psi_f(\eta) = \int_0^\eta \varphi(\lambda) d\lambda.$$

Figure 1:



Theorem 3.2: *Let \mathbb{C} be a n -deviation measurable (v,k) -constructible VC-class. Suppose that $\sup f(x) < \infty$ and that f has at most finitely many critical levels. If $f \in F_{\mathbb{C}}$ then we have*

$$d_1(f_{n,\mathbb{C}}, f)^* = O_p(\Psi(n^{-1/3} (\log n)^{1/3})) \quad \text{as } n \rightarrow \infty.$$

Remarks 3.3: (i) If the support of f has finite Lebesgue measure, then Theorem 3.2 gives the rate $n^{-1/3} (\log n)^{1/3}$, because in this case $\Psi(\eta) = O(\eta)$ as $\eta \rightarrow 0$. If the support of f has infinite Lebesgue measure, then $\Psi(\eta)$ tends to zero (as $\eta \rightarrow 0$) slower than $O(\eta)$. This leads to slower rates of convergence of $f_{n,\mathbb{C}}$. For example for the normal distribution in \mathbf{R}^d we have $\Psi(\eta) = O(\eta (\log 1/\eta)^{d/2})$.

(ii) As already mentioned, \mathbb{C}_0 is a VC-class. Hence Theorem 3.2 also gives a rate of convergence for the Grenander estimator: If f has a bounded support, then this rate is $n^{-1/3} (\log n)^{1/3}$ (Groeneboom [6] showed that $n^{-1/3}$ is the exact rate). Hence, although we did not use any special properties of the Grenander estimator (such as monotonicity for example) we derived the exact rate up to a log-term by only using the fact that the corresponding class \mathbb{C}_0 is a VC-class.

In the following theorem we also allow more richer classes than VC-classes. The richness is measured in terms of the metric entropy with inclusion of \mathbb{C} with respect to F , which is defined as follows. Let

$$N_I(\varepsilon, \mathbb{C}, F) := \inf \left\{ m \in \mathbf{N}: \exists C_1, \dots, C_m \text{ measurable, such that for every } C \in \mathbb{C} \text{ there exist } i, j \in \{1, \dots, m\} \text{ with } C_i \subset C \subset C_j \text{ and } F(C_j \setminus C_i) < \varepsilon \right\},$$

then $\log N_I(\varepsilon, \mathbb{C}, F)$ is called metric entropy with inclusion of \mathbb{C} with respect to F .

Theorem 3.4: *Let \mathbb{C} be such that there exist constants $A, r > 0$ with*

$$(3.3) \quad \log N_I(\varepsilon, \mathbb{C}, F) \leq A \varepsilon^{-r} \quad \forall \varepsilon > 0.$$

Suppose that $\sup f(x) < \infty$ and that f has at most finitely many critical levels. Then we have

$$d_1(f_{n,\mathbb{C}}, f)^* = O_p(\Psi(\alpha_n)) \quad \text{as } n \rightarrow \infty,$$

where

$$\alpha_n = \begin{cases} n^{-1/(3+r)}, & r < 1 \\ n^{-1/4} \log(n), & r = 1 \\ n^{-1/2(r+1)}, & r > 1. \end{cases}$$

Examples: Let $\mathbb{C} = \mathbb{E}^d$, $d \geq 2$. If the support of f is compact and $\sup\{f(x)\} < \infty$, then (3.3) is satisfied with $r = (d-1)/2$ (see e.g. Dudley [2]). Hence it follows from Theorem 3.4 that in this case

$$d_I(f_{n, \mathbb{E}^d}, f)^* = O_{P^*} \left(\begin{array}{ll} n^{-2/7}, & d = 2 \\ n^{-1/4} \log n, & d = 3 \\ n^{-1/(d+1)}, & d \geq 4 \end{array} \right).$$

Under mild conditions on the tail behaviour of f the assumption of a compact support can be dropped (see Polonik [10]), so that for example these rates also hold for the normal distribution up to an additional log-term which comes in through the behaviour of Ψ (see Remark 3.3 (i)).

The case of an underlying uniform distribution:

Theorem 3.2 and Theorem 3.4 can of course also be applied to an underlying uniform distribution. However, there they lead to rates which are far from the optimal. This can be seen in the special case of the Grenander estimator. The rate of convergence of the Grenander estimator is known to be $n^{-1/2}$ in the case of an underlying uniform distribution. However, Theorem 3.2 only gives the upper bound $n^{-1/3} (\log n)^{1/3}$. But we are able to re-derive the correct rate $n^{-1/2}$ up to a log-term (see Corollary 3.5 (a) below).

Theorem 3.5: *Let F be a uniform distribution on a set C with $\text{Leb}(C) < \infty$. Let $\{\beta_n\}$ be a sequence of real numbers converging to zero as $n \rightarrow \infty$. Suppose that $\|F_n - F\|_{\mathbb{C}} = O_P(\beta_n)$ as $n \rightarrow \infty$. If $C \in \mathbb{C}$ then we have*

$$d_I(f_{n, \mathbb{C}}, f)^* = O_P(\beta_n \log(1/\beta_n)) \quad \text{as } n \rightarrow \infty.$$

Corollary 3.6: Let F be a uniform distribution on a set C with $\text{Leb}(C) < \infty$ and suppose that $C \in \mathbb{C}$.

(a) If \mathbb{C} is an n -deviation measurable (v,m) -constructible VC-class then we have

$$d_1(f_{n,\mathbb{C}}f)^* = O_p(n^{-1/2} \log n) \quad \text{as } n \rightarrow \infty.$$

(b) If \mathbb{C} satisfies the entropy condition (3.3) then we have

$$d_1(f_{n,\mathbb{C}}f)^* = O_p(\alpha_n), \quad \text{as } n \rightarrow \infty$$

where

$$\alpha_n = \begin{cases} n^{-1/2} \log(n), & r < 1 \\ n^{-1/2}(\log n)^2, & r = 1 \\ n^{-1/(r+1)}\log(n), & r > 1. \end{cases}$$

4. Concluding Remarks

◆ *Computation:* In the one-dimensional case, more precisely, for the case $\mathbb{C} = \mathbf{I}_k$, $k = 1, 2, 3$ there exists a computer program of Müller and Sawitzki [8]) for calculating the empirical generalized λ -clusters. Since empirical generalized λ -clusters are minimum volume sets (however, for a *random* parameter α ; cf. introduction and Section 2), it is possible to use algorithms for calculating minimum volume sets in order to calculate empirical generalized λ -clusters. In higher dimensions Nolan [9] gave some calculations for the case $\mathbb{C} = \mathbf{E}^d$. One first has to calculate *all* minimum volume ellipsoids (MVE), i.e. the MVE for all parameters $\alpha_n = k/n$, $k = 1, \dots, n$. Then in a second step it is easy to calculate the empirical generalized λ -clusters in \mathbf{E}^d for all $\lambda \geq 0$. For the case $\mathbb{C} = \mathbf{E}^2$, Hartigan [4] gave an algorithm for directly calculating a set $\Gamma_{n,\mathbf{E}^2}(\lambda)$ for a fixed $\lambda \geq 0$. However, for classes \mathbb{C} of sets with more complicated shapes there exist no algorithms until now.

◆ *Relaxing the assumptions on critical values:* Instead of assuming that f has at most finitely many flat parts (see Theorem 3.2 and 3.4) one can make the following weaker assumption: *for each $\eta > 0$ there exists a constant $C_{\lambda,\eta} < \infty$ and a region $\Lambda_\eta \subset [0, \infty)$ with $[0, \max(f(x))] \setminus \Lambda_\eta = O(\eta)$ such that*

$$\text{Leb}\{x: |f(x) - \lambda| < \eta\} \leq C_{\lambda,\eta} \eta, \quad \forall \lambda \in \Lambda_\eta,$$

and that for each $\eta > 0$, $C_{\lambda, \eta}^{-1}$ is integrable over Λ_η

◆ *The case where $f \notin F_{\mathbb{C}}$* : This case can be interpreted as a wrong model. Let the sets $\Gamma_{\mathbb{C}}(\lambda)$ be defined as those sets which maximize the measure $H_\lambda = F - \lambda \text{Leb}$ over the class \mathbb{C} , i.e.

$$H_\lambda(\Gamma_{\mathbb{C}}(\lambda)) = \sup_{C \in \mathbb{C}} H_\lambda(C).$$

This definition is completely analogous to the definition of $\Gamma_{n, \mathbb{C}}(\lambda)$. The sets $\Gamma_{\mathbb{C}}(\lambda)$ are called *generalized λ -clusters*. (Suppose that they exist for each $\lambda > 0$ and that they are unique (up to F -nullsets)). It has been shown in Polonik [10] that for normalizing classes $d_1(f_{n, \mathbb{C}}, f_{\mathbb{C}})^*$ converges to zero with probability 1, where

$$f_{\mathbb{C}}(x) = \int \mathbb{I}_{\Gamma_{\mathbb{C}}(\lambda)}(x) d\lambda \quad \forall x \in \mathbf{R}.$$

Here \mathbb{C} has to satisfy some additional assumptions (which all are satisfied for the standard class $\mathbf{1}_1, \mathbf{B}^d, \mathbf{E}^d$ and \mathbf{C}^d). It is difficult to interpret the function $f_{\mathbb{C}}$. However, it can be shown, that $f_{\mathbb{C}}$ often is a probability density. More precisely, one can show, that the integral of $f_{\mathbb{C}}$ over \mathbf{R}^d equals $\lim_{\delta \rightarrow 0} F(\Gamma_{\mathbb{C}}(\delta))$ (see Polonik [10]). Often this limit equals $F(\Gamma_{\mathbb{C}}(0))$, and for standard classes \mathbb{C} we often have $F(\Gamma_{\mathbb{C}}(0)) = 1$.

◆ *Uniform rates of $f_{n, \mathbb{C}}$* : The rates given in the Theorems 3.2 and 3.4 also hold uniformly over a class of densities F if the tail behaviour in this class can be controlled uniformly. More precisely, let $\sup_{f \in F} \Psi_f(\eta) = \tilde{\Psi}_F(\eta)$, then we have

Theorem 4.1: *Let F be a class of densities such that every $f \in F$ has at most finitely many critical levels. Let \mathbb{C} be such that (3.3) is satisfied. Then we have that for every $\varepsilon > 0$ there exists a constant $c > 0$ such that*

$$\sup_{f \in F} P [d_1(f_{n, \mathbb{C}}, f)^* \geq c \tilde{\Psi}_F(\alpha_n)] \leq \varepsilon \quad \text{for all } n \geq n_0(\varepsilon)$$

where α_n is the same as in Theorem 3.4. If \mathbb{C} is an n -deviation measurable VC-class then the same result holds with $\alpha_n = n^{-1/3} (\log n)^{1/3}$.

The proof of Theorem 4.1 is the same as the proofs of Theorem 3.4 and Theorem 3.2, respectively. One only has to notice, that Proposition 5.2 actually holds uniformly over F (this can be seen from the proof).

◆ *Remark on assumption (γ) :* For many standard classes \mathbb{C} we have $F_n(\Gamma_{n,\mathbb{C}}(\lambda)) = O(n^{-1})$ a.s. for $\lambda > \lambda_{n,\max}$ as $n \rightarrow \infty$. As in Polonik [10] this (weaker) property could be assumed to hold instead of (γ) . However, in this case an additional $O(n^{-1})$ would enter the calculations. Therefore, in order to shorten the formulas, we prefer to assume (γ) .

5. Proofs

Proof of Proposition 2.2: Note that

$$\begin{aligned} E_{n,\mathbb{C}}(\lambda) &= \sup \{ (F_n - \lambda \text{Leb})(C) : C \in \mathbb{C} \} \\ &= \max_{j \in \{0, \dots, n\}} [j/n - \lambda \inf_{\mathbb{C}(j) \neq \emptyset} \{ \text{Leb}(C) : C \in \mathbb{C} \}]. \end{aligned}$$

Hence $E_{n,\mathbb{C}}$ equals the maximum over at most $n + 1$ (different) linear functions with slope ≤ 0 and is therefore a monotone decreasing concave function with at most $n+1$ changes of slope. Choose $0 \leq \lambda_0 < \lambda_1 < \dots < \lambda_{k_n}$, $k_n \leq n$ as those values of λ where the slope of $E_{n,\mathbb{C}}$ changes. It follows from the above representation of $E_{n,\mathbb{C}}$ that for all $\lambda \notin \{\lambda_0, \lambda_1, \dots, \lambda_{k_n}\}$ $\text{Leb}(\Gamma_{n,\mathbb{C}}(\lambda))$ equals the slope of $E_{n,\mathbb{C}}$ at λ . Hence, for these values of λ the monotonicity of $\text{Leb}(\Gamma_{n,\mathbb{C}}(\lambda))$ follows. If however the monotonicity would be violated for a value $\lambda \in \{\lambda_0, \lambda_1, \dots, \lambda_{k_n}\}$ then the above representation of $E_{n,\mathbb{C}}$ would easily give a contradiction. This proves the first part of (a). Now it again follows from the above representation of $E_{n,\mathbb{C}}$ that every set $\Gamma_{n,\mathbb{C}}(\lambda_j)$, $j = 0, \dots, k_n - 1$ is an empirical generalized λ -cluster for all $\lambda \in [\lambda_j, \lambda_{j+1}]$. This proves the second part of (a).

In order to see (b) first note that $\Gamma_{n,\mathbb{C}}(\lambda_{k_n})$ is an empirical generalized λ -cluster for all $\lambda \geq \lambda_{k_n}$. Hence we have for every $\lambda \geq \lambda_{k_n}$ that $E_{n,\mathbb{C}}(\lambda) = F_n(\Gamma_{n,\mathbb{C}}(\lambda_{k_n})) - \lambda \text{Leb}(\Gamma_{n,\mathbb{C}}(\lambda_{k_n}))$. Since $\emptyset \in \mathbb{C}$ we have $E_{n,\mathbb{C}}(\lambda) \geq 0$ and therefore $\text{Leb}(\Gamma_{n,\mathbb{C}}(\lambda_{k_n})) = 0$. Otherwise $E_{n,\mathbb{C}}(\lambda)$ would be strictly smaller than zero for λ large enough which would give a contradiction. □

Proof of Proposition 2.3: Let us first assume that the empirical generalized λ -clusters are

monotone for inclusion. Below we shall show that the general situation can be reduced to this “case of monotonicity” via symmetrization arguments.

Let λ_j , $j = 1, \dots, k_n$ be the (random) levels of assumption (α) (see Section 2, or proof of Proposition 2.2 above). Note that the assumption $\lambda_{n, \max} > 0$ is equivalent to $k_n \geq 1$. For every $j \in 0, \dots, k_n - 1$ we define

$$\Delta_{n, \mathbb{C}}(j) := \Gamma_{n, \mathbb{C}}(\lambda_j) \setminus \Gamma_{n, \mathbb{C}}(\lambda_{j+1}).$$

The monotonicity assumption says that $\Gamma_{n, \mathbb{C}}(\lambda_{j+1}) \subsetneq \Gamma_{n, \mathbb{C}}(\lambda_j)$, $j = 0, \dots, k_n - 1$. Thus the sets $\Delta_{n, \mathbb{C}}(j)$, $j = 0, \dots, k_n - 1$, are disjoint and we have

$$\bigcup_{j=1}^{k_n-1} \Delta_{n, \mathbb{C}}(j) = \Gamma_{n, \mathbb{C}}(0) \setminus \Gamma_{n, \mathbb{C}}(\lambda_{k_n}).$$

Furthermore it follows together with (2.3) that $f_{n, \mathbb{C}}$ is constant on $\Delta_{n, \mathbb{C}}(j)$ with $f_{n, \mathbb{C}}(x) = \lambda_{j+1}$ for all $x \in \Delta_{n, \mathbb{C}}(j)$. Hence we have

$$(5.1) \quad \int f_{n, \mathbb{C}}(x) \, dx = \sum_{j=0}^{k_n-1} \int_{\Delta_{n, \mathbb{C}}(j)} f_{n, \mathbb{C}}(x) \, dx = \sum_{j=0}^{k_n-1} \lambda_{j+1} \text{Leb}(\Delta_{n, \mathbb{C}}(j)).$$

Now we derive a representation for λ_{j+1} . Together with (5.1) this representation will give the assertion. Note that for every $j \in \{0, \dots, k_n - 1\}$ the empirical generalized λ_j -cluster also is an empirical generalized λ_{j+1} -cluster (see proof of Proposition 2.2). Hence it follows that for every $j \in \{0, \dots, k_n - 1\}$ we have

$$F_n(\Gamma_{n, \mathbb{C}}(\lambda_{j+1})) - \lambda_{j+1, n} \text{Leb}(\Gamma_{n, \mathbb{C}}(\lambda_{j+1})) = F_n(\Gamma_{n, \mathbb{C}}(\lambda_j)) - \lambda_{j+1} \text{Leb}(\Gamma_{n, \mathbb{C}}(\lambda_j)).$$

From this equality we obtain that

$$(5.2) \quad \begin{aligned} \lambda_{j+1} &= [F_n(\Gamma_{n, \mathbb{C}}(\lambda_{j+1})) - F_n(\Gamma_{n, \mathbb{C}}(\lambda_j))] / [\text{Leb}(\Gamma_{n, \mathbb{C}}(\lambda_{j+1})) - \text{Leb}(\Gamma_{n, \mathbb{C}}(\lambda_j))], \\ &= F_n(\Delta_{n, \mathbb{C}}(j)) / \text{Leb}(\Delta_{n, \mathbb{C}}(j)). \end{aligned}$$

From (5.1) and (5.2) the assertion follows:

$$\int f_{n,\mathbb{C}}(x) dx = \sum_{j=0}^{k_n-1} F_n(\Delta_{n,\mathbb{C}}(j)) = F_n(\Gamma_{n,\mathbb{C}}(0)) - F_n(\Gamma_{n,\mathbb{C}}(\lambda_{k_n})).$$

Now we show that the general case, i.e. the case without the monotonicity assumption, can be reduced to the special case proven above.

As already mentioned, $f_{n,\mathbb{C}}(x)$ can be constructed in the “case of monotonicity” by putting the slices $S_{n,\mathbb{C}}(\lambda_j) := \Gamma_{n,\mathbb{C}}(\lambda_{j-1}) \times [\lambda_j, \lambda_{j-1}]$, $j = 1, \dots, k_n$ one on the top of the other. However, if we do this and the sets $\Gamma_{n,\mathbb{C}}(\lambda_j)$ are not monotone, then the resulting figure in \mathbf{R}^{d+1} does not equal the subgraph of $f_{n,\mathbb{C}}(x)$, i.e. $S_{n,\mathbb{C}} := \bigcup_{j=0}^{k_n} S_{n,\mathbb{C}}(\lambda_j) \neq \{(x,y) \in \mathbf{R}^d \times \mathbf{R} : 0 \leq y \leq f_{n,\mathbb{C}}(x), x \in \Gamma_{n,\mathbb{C}}(0)\}$. Nevertheless, in any case, the volume of $S_{n,\mathbb{C}}$ in \mathbf{R}^{d+1} equals the L_1 -norm of $f_{n,\mathbb{C}}(x)$.

The volume of $S_{n,\mathbb{C}}$ does not change if some of the $S_{n,\mathbb{C}}(\lambda_j)$ are replaced by sets with the same volume. We shall replace the sets $\Gamma_{n,\mathbb{C}}(\lambda)$ by their so-called “Schwarz symmetrizations” $\tilde{\Gamma}_{n,\mathbb{C}}(\lambda)$. They are defined as balls with midpoint zero which have the same Lebesgue measure as $\Gamma_{n,\mathbb{C}}(\lambda)$. The sets $\tilde{\Gamma}_{n,\mathbb{C}}(\lambda)$ are monotonically decreasing (in λ) for inclusion, because the Lebesgue measures of the empirical λ -clusters are decreasing in λ (Proposition 2.2). Hence, replacing the sets $\Gamma_{n,\mathbb{C}}(\lambda)$ by $\tilde{\Gamma}_{n,\mathbb{C}}(\lambda)$ and putting the resulting symmetrized slices $\tilde{\Gamma}_{n,\mathbb{C}}(\lambda_{j-1,n}) \times [\lambda_j, \lambda_{j-1}]$, $j = 1, \dots, k_n$ one on the top of the other, gives us a function $\tilde{f}_{n,\mathbb{C}}$ which has the same properties as $f_{n,\mathbb{C}}$ used in the above proven special case: It is a pure jump function which takes *the same constant values* λ_j , $j = 0, \dots, k_n - 1$ as $f_{n,\mathbb{C}}$ on sets which have the same Lebesgue measure as the sets $\Delta_{n,\mathbb{C}}(j)$. By construction $\tilde{f}_{n,\mathbb{C}}$ has the same L_1 -norm as $f_{n,\mathbb{C}}$. This L_1 -norm can be calculated as in the “case of monotonicity” treated above. □

Proof of Theorem 3.1: As mentioned already, we shall use results about the empirical generalized λ -clusters. One of these results is the following inequality which has been used in Polonik [10, 11]. (For completeness we give the proof of the inequality below.) If $\Gamma(\lambda) \in \mathbb{C}$ then we have for every $\eta > 0$

$$(5.3) \quad d_j(\Gamma(\lambda), \Gamma_{n,\mathbb{C}}(\lambda)) \leq \text{Leb}\{x: |f(x) - \lambda| < \eta\} + \eta^{-1} [(F_n - F)(\Gamma_{n,\mathbb{C}}(\lambda)) - (F_n - F)(\Gamma(\lambda))].$$

For any measurable set $\Lambda \subset [0, \infty)$ define

$$f_{n, \mathbb{C}}(x, \Lambda) := \int_{\Lambda} \mathbb{1}_{\Gamma_{n, \mathbb{C}}(\lambda)}(x) d\lambda.$$

Analogously we define $f(x, \Lambda)$ with $\Gamma(\lambda)$ instead of $\Gamma_{n, \mathbb{C}}(\lambda)$. We need the following lemma which will be proven below:

Lemma 5.1: *For every measurable set $\Lambda \subset [0, \infty)$ and every normalizing class \mathbb{C} we have*

$$d_I(f_{n, \mathbb{C}}, f) \leq 2 \int_{\Lambda} d_I(\Gamma_{n, \mathbb{C}}(\lambda), \Gamma(\lambda)) d\lambda + 2 \int f(x, \Lambda^c) dx$$

where $\Lambda^c = [0, \infty) \setminus \Lambda$.

For $0 < \eta < L$ let $\Lambda_{\eta, L} := [\eta, L]$. It follows from Lemma 5.1 that

$$(5.4) \quad d_I(f_{n, \mathbb{C}}, f) \leq 2 \int_{\eta}^L d_I(\Gamma_{n, \mathbb{C}}(\lambda), \Gamma(\lambda)) d\lambda + 2 \int f(x, \Lambda_{\eta, L}^c) dx.$$

Let $\varphi(\lambda) := \text{Leb}(\Gamma(\lambda)) = \text{Leb}\{x : f(x) \geq \lambda\}$, $\lambda \geq 0$. The second integral on the right-hand side in (5.4) can be written as $2 \int_M^{\infty} \varphi(\lambda) d\lambda + 2 \int_0^{\eta} \varphi(\lambda) d\lambda$. Both integrals can be made arbitrarily small by choosing L large enough and η small enough, respectively. (Note that $\int_L^{\infty} \varphi(\lambda) d\lambda = \int (f(x) - L)^+ dx$ and $\int_0^{\eta} \varphi(\lambda) d\lambda = \int (\eta \wedge f(x)) dx$). As for the first integral on the right-hand side in (5.4) it follows from (5.3) that

$$\begin{aligned} \int_{\eta}^L d_I(\Gamma_{n, \mathbb{C}}(\lambda), \Gamma(\lambda)) d\lambda &\leq \int_{\eta}^L \text{Leb}\{x : |f(x) - \lambda| < \eta\} d\lambda \\ &\quad + \eta^{-1} \int_{\eta}^L [(F_n - F)(\Gamma_{n, \mathbb{C}}(\lambda)) - (F_n - F)(\Gamma(\lambda))] d\lambda. \end{aligned}$$

Hence we have

$$d_I(f_{n, \mathbb{C}}, f) \leq 2 \eta^{-1} \int_{\eta}^L [(F_n - F)(\Gamma_{n, \mathbb{C}}(\lambda)) - (F_n - F)(\Gamma(\lambda))] d\lambda + A(\eta, L)$$

with

$$A(\eta, L) := 2 \int f(x, (\Lambda_{\eta, L})^c) dx + 2 \int_{\eta}^L \text{Leb}\{x: |f(x) - \lambda| < \eta\} d\lambda.$$

Now we show that $A(\eta, L) \rightarrow 0$ as $\eta \rightarrow 0$ and $L \rightarrow \infty$. We already know (see above) that the first integral in the expression of $A(\eta, L)$ converges to zero as $\eta \rightarrow 0$ and $L \rightarrow \infty$, respectively. Hence it remains to consider the second integral. Note that for η small enough we have

$$(5.5) \quad \text{Leb}\{x: |f(x) - \lambda| < \eta\} = \varphi(\lambda - \eta) - \varphi(\lambda + \eta) - \text{Leb}\{x: f(x) = \lambda - \eta\}.$$

Since there exist at most countable many levels μ with $\text{Leb}\{x: f(x) = \mu\} \neq 0$ we have

$$\begin{aligned} \int_{\eta}^L \text{Leb}\{x: |f(x) - \lambda| < \eta\} d\lambda &= \int_{\eta}^L \varphi(\lambda - \eta) - \varphi(\lambda + \eta) d\lambda \\ &= \int_0^{2\eta} \varphi(\lambda) d\lambda - \int_{L-\eta}^{L+\eta} \varphi(\lambda) d\lambda. \end{aligned}$$

Since f is integrable, both integrals in the last line converge to zero as $\eta \rightarrow 0$, and therefore $A(\eta, L) \rightarrow 0$ as $\eta \rightarrow 0$ and $L \rightarrow \infty$. To finish the proof of the theorem it remains to show that the measurable cover of

$$\int_{\eta}^L [(F_n - F)(\Gamma_{n, \mathcal{C}}(\lambda)) - (F_n - F)(\Gamma(\lambda))] d\lambda$$

converges to zero with probability 1 as $n \rightarrow \infty$ for every fixed η and L with $L > \eta > 0$. But this follows from the GC(F)-property of \mathbb{C} . □

Proof of Theorem 3.2 and Theorem 3.4: Let α_n be as follows:

$$\alpha_n = \begin{cases} n^{-1/3} (\log n)^{1/3} & \text{if } \mathbb{C} \text{ is a VC-class} \\ \text{defined as in Theorem 3.4 if } \mathbb{C} \text{ satisfies (3.3).} & \end{cases}$$

First we give an outline of the proof in some heuristic arguments. In (3.2) we showed that

$$(5.6) \quad d_I(f_{n,\mathbb{C}}, f) \leq \int d_I(\Gamma_{n,\mathbb{C}}(\lambda), \Gamma(\lambda)) d\lambda.$$

Therefore we look at the behaviour of $d_I(\Gamma_{n,\mathbb{C}}(\lambda), \Gamma(\lambda))$. If φ is differentiable at λ then $\text{Leb}\{x : |f(x) - \lambda| < \eta\} = O(\eta)$ (cf. (5.5)), and one can show (see Polonik [11]) that in this situation $d_I(\Gamma_{n,\mathbb{C}}(\lambda), \Gamma(\lambda)) = O_P(\alpha_n)$. However, in general these rates do not hold uniform over λ . But we shall show (Proposition 5.2 below), that there exists a function $g_n(\lambda)$ such that for a “large enough”(see below) region $\Lambda_n \subset [0, \infty)$ we have

$$\sup_{\lambda \in \Lambda_n} [g_n(\lambda) d_I(\Gamma_{n,\mathbb{C}}(\lambda), \Gamma(\lambda))]^* = O_P(\alpha_n).$$

If in addition $g_n^{-1}(\lambda)$ is integrable over Λ_n , then it follows that

$$d_I(f_{n,\mathbb{C}}, f)^* = O_P\left(\alpha_n \int_{\Lambda_n} g_n(\lambda)^{-1} d\lambda\right).$$

It will turn out that $\alpha_n \int_{\Lambda_n} g_n(\lambda)^{-1} d\lambda = \Psi(\alpha_n)$, such that the assertions of Theorem 3.2 and Theorem 3.4, respectively, follow. Note that in general $d_I(\Gamma_{n,\mathbb{C}}(\lambda), \Gamma(\lambda))$ need not converge to zero at critical levels λ (cf. Section 3). Therefore, in (5.6), we shall leave out “small” neighbourhoods around critical levels, such that these neighbourhoods tend to zero fast enough as n tends to infinity. This leads to the “large enough” region Λ_n considered above.

As above let $\varphi(\lambda) := \text{Leb}(\Gamma(\lambda)) = \text{Leb}\{x : f(x) \geq \lambda\}$ and let φ' denote the derivative of φ with respect to λ . The proof of the next proposition will be given at the end of Section 5.

Proposition 5.2: *Suppose that the assumptions of Theorem 3.2 and Theorem 3.4, respectively, hold. Let α_n be as above and let $a = a_n$ and $b = b_n$, $0 < a < b \leq \infty$, such that the interval $(a - \alpha_n, b + \alpha_n]$ contains no critical level. For $\lambda \in (a, b)$ let $\xi_{\lambda,n}$ be defined through the equation*

$$\text{Leb}\{x : |f(x) - \lambda| < \alpha_n\} = \varphi(\lambda - \alpha_n) - \varphi(\lambda + \alpha_n) = -2 \varphi'(\xi_{\lambda,n}) \alpha_n.$$

Then we have as $n \rightarrow \infty$

$$\sup_{\lambda \in (a,b)} [|\varphi'(\xi_{\lambda,n})| \vee 1]^{-1} d_I(\Gamma_{n,\mathbb{C}}(\lambda), \Gamma(\lambda))]^* = O_P(\alpha_n).$$

Let $M = \sup f(x) < \infty$. We obtain from Lemma 5.1 with $\Lambda = (\alpha_n, M)$ that

$$d_I(f_{n, \mathbb{C}}, f) \leq 2 \int_{\alpha_n}^M d_I(\Gamma_{n, \mathbb{C}}(\lambda), \Gamma(\lambda)) + 2 \int_0^{\alpha_n} \varphi(\lambda) d\lambda$$

Now we consider the first integral on the right-hand side. To simplify the notation we assume that f has only one critical level $\lambda_0 > 0$. (The proof for the case of more than one critical level is completely analogous). Suppose that $\lambda_0 < M$, then we have

$$\begin{aligned} \int_{\alpha_n}^M d_I(\Gamma_{n, \mathbb{C}}(\lambda), \Gamma(\lambda)) d\lambda &= \int_{\alpha_n}^{\lambda_0 - \alpha_n} d_I(\Gamma_{n, \mathbb{C}}(\lambda), \Gamma(\lambda)) d\lambda \\ &+ \int_{\lambda_0 - \alpha_n}^{\lambda_0 + \alpha_n} d_I(\Gamma_{n, \mathbb{C}}(\lambda), \Gamma(\lambda)) d\lambda \\ (5.7) \quad &+ \int_{\lambda_0 + \alpha_n}^M d_I(\Gamma_{n, \mathbb{C}}(\lambda), \Gamma(\lambda)) d\lambda. \end{aligned}$$

The second integral on the right-hand side of (5.7) is of the order $O_p(\alpha_n)$, because for any fixed $\varepsilon > 0$ we have for large enough n (such that $\alpha_n < \varepsilon$) that

$$\begin{aligned} \sup_{\lambda \in (\lambda_0 - \alpha_n, \lambda_0 + \alpha_n)} d_I(\Gamma_{n, \mathbb{C}}(\lambda), \Gamma(\lambda)) &\leq \sup_{\lambda \in (\lambda_0 - \alpha_n, \lambda_0 + \alpha_n)} [Leb(\Gamma_{n, \mathbb{C}}(\lambda)) + Leb(\Gamma(\lambda))] \\ &\leq Leb(\Gamma_{n, \mathbb{C}}(\lambda_0 - \varepsilon)) + Leb(\Gamma(\lambda_0 - \varepsilon)) \\ &\leq 2 Leb(\Gamma(\lambda_0 - \varepsilon)) + O_p(1) = O_p(1). \end{aligned}$$

The second inequality of this chain of inequalities follows from the monotonicity of the functions $\lambda \rightarrow Leb(\Gamma(\lambda))$ and $\lambda \rightarrow Leb(\Gamma_{n, \mathbb{C}}(\lambda))$ (for the latter see Proposition 2.2) and the third inequality follows from the consistency of $\Gamma_{n, \mathbb{C}}(\lambda)$ for all λ that are no critical level (Polonik [10, 11]).

Now we consider the first and the third integral on the right-hand side of (5.7). They both are of the form $\int_a^b d_I(\Gamma_{n, \mathbb{C}}(\lambda), \Gamma(\lambda)) d\lambda$, where a and b fulfill the requirements of Proposition 5.2. Hence it follows from Proposition 5.2 and the definition of $\xi_{\lambda, n}$ that as $n \rightarrow \infty$

$$\begin{aligned}
\int_a^b d_I(\Gamma_{n,C}(\lambda), \Gamma(\lambda)) d\lambda &\leq O_{P^*}(\alpha_n) \int_a^b [|\varphi'(\xi_{\lambda,n})| \vee 1] d\lambda \\
&\leq O_{P^*}(\alpha_n) [(b + \alpha_n - a) + \int_a^b |\varphi'(\xi_{\lambda,n})| d\lambda] \\
&= O_{P^*}(\alpha_n) + O_{P^*}(\int_a^b \varphi(\lambda - \alpha_n) - \varphi(\lambda + \alpha_n) d\lambda) \\
&= O_{P^*}(\alpha_n) + O_{P^*}(\int_{a-\alpha_n}^{a+\alpha_n} \varphi(\lambda) d\lambda - \int_{b-\alpha_n}^{b+\alpha_n} \varphi(\lambda) d\lambda) \\
&\leq O_{P^*}(\alpha_n) + O_{P^*}(\int_{a-\alpha_n}^{a+\alpha_n} \varphi(\lambda) d\lambda)
\end{aligned}$$

The last inequality holds since φ is decreasing. Now we apply these upper bounds to the first and the third integral on the right-hand side of (5.7). This gives

$$\int_{\lambda_0 + \alpha_n}^M d_I(\Gamma_{n,C}(\lambda), \Gamma(\lambda)) d\lambda \leq O_{P^*}(\alpha_n) + O_{P^*}(\int_{\lambda_0}^{\lambda_0 + 2\alpha_n} \varphi(\lambda) d\lambda) = O_{P^*}(\alpha_n),$$

and

$$\int_{\alpha_n}^{\lambda_0 - \alpha_n} d_I(\Gamma_{n,C}(\lambda), \Gamma(\lambda)) d\lambda \leq O_{P^*}(\alpha_n) + O_{P^*}(\int_0^{2\alpha_n} \varphi(\lambda) d\lambda) = O_{P^*}(\Psi(\alpha_n)).$$

The assertion of the theorems follow by collecting the just derived upper bounds of the three integrals on the right-hand side of (5.7).

For $\lambda_0 = M$ we of course split the integral on the left-hand side into two integrals extended over $(\alpha_n, \lambda_0 - \alpha_n)$ and $(\lambda_0 - \alpha_n, \lambda_0)$, respectively. Upper bounds for these two integrals can be obtained analogously to the one given above. They lead to the same results.

□

Proof of Theorem 3.5: Let $M = M(C) = 1/\text{Leb}(C)$. First we show that

$$(5.8) \quad [M - \lambda] [\text{Leb}(C \Delta \Gamma_{n,C}(\lambda))] \leq 2 \sup_{C \in \mathcal{C}} |F_n(C) - F(C)|.$$

This can be seen as follows:

$$\begin{aligned} H_\lambda(\Gamma_C(\lambda)) - H_\lambda(\Gamma_{n,C}(\lambda)) &= [F(C) - \lambda \text{Leb}(C)] - [F(\Gamma_{n,C}(\lambda)) - \lambda \text{Leb}(\Gamma_{n,C}(\lambda))] \\ &= [1 - \lambda \text{Leb}(C)] - [\text{Leb}(\Gamma_{n,C}(\lambda)) / \text{Leb}(\Gamma_{n,C}(\lambda)) - \lambda \text{Leb}(\Gamma_{n,C}(\lambda))] \\ &= [1 / \text{Leb}(C) - \lambda] [\text{Leb}(C) - \text{Leb}(\Gamma_{n,C}(\lambda))] \\ &= [M - \lambda] [\text{Leb}(C \Delta \Gamma_{n,C}(\lambda))]. \end{aligned}$$

Hence

$$[M - \lambda] [\text{Leb}(C \Delta \Gamma_{n,C}(\lambda))] \leq (F_n - F)(\Gamma_{n,C}(\lambda)) - (F_n - F)(C).$$

This proves (5.8). An application of Lemma 5.1 with $\Lambda = [0, M]$ together with (5.8) gives

$$\begin{aligned} d_1(f_{n,C}, f) &\leq 2 \int_0^M \text{Leb}(\Gamma_{n,C}(\lambda) \Delta C) d\lambda \\ &= 2 \left[\int_0^{M-\beta_n} + \int_{M-\beta_n}^M \right] \text{Leb}(\Gamma_{n,C}(\lambda) \Delta C) d\lambda \\ &= O_p(\beta_n) \int_0^{M-\beta_n} (M(C) - \lambda)^{-1} d\lambda + \int_{M-\beta_n}^M \text{Leb}(\Gamma_{n,C}(\lambda) \Delta C) d\lambda \end{aligned}$$

Now, the first term in the last line is of the order $O_p(\beta_n) O(\log 1/\beta_n)$. It remains to show that the second integral in the last line is not of slower order. Actually we have

$$\int_{M-\beta_n}^M \text{Leb}(\Gamma_{n,C}(\lambda) \Delta C) d\lambda = O_p(\beta_n).$$

This follows from $\sup_{M-\beta_n < \lambda < M} \text{Leb}(\Gamma_{n,C}(\lambda) \Delta C) = O_p(1)$, which in turn can be proven by analogous arguments as given after (5.7) above. □

Proof of Corollary 3.6: By taking $\alpha = 1/2$ it follows from Theorem 5.3 that in the situation of (a), $\|F_n - F\|_{\mathbb{C}} = O_{\mathbb{P}}(n^{-1/2})$, and in the situation given in (b) we have

$$\|F_n - F\|_{\mathbb{C}} = O_{\mathbb{P}} \left(\begin{cases} n^{-1/2}, & r < 1 \\ n^{-1/2} \log n, & r = 1 \\ n^{-1/(r+1)}, & r > 1 \end{cases} \right).$$

Now the assertions follow directly from Theorem 3.5. □

Proofs of Section 5:

Proof of inequality (5.3): First note that

$$\begin{aligned} H_{\lambda}(\Gamma(\lambda)) - H_{\lambda}(\mathbb{C}) &= \int_{\Gamma(\lambda)} (f(x) - \lambda) dx - \int_{\mathbb{C}} (f(x) - \lambda) dx \\ &= \int_{\Gamma(\lambda) \setminus \mathbb{C}} (f(x) - \lambda) dx - \int_{\mathbb{C} \setminus \Gamma(\lambda)} (f(x) - \lambda) dx \\ &= \int_{\Gamma(\lambda) \Delta \mathbb{C}} |f(x) - \lambda| dx. \end{aligned}$$

To shorten the notation we denote $D_{n,\mathbb{C}}(\lambda) = \Gamma_{n,\mathbb{C}}(\lambda) \Delta \Gamma_{\mathbb{C}}(\lambda)$, so that $\text{Leb}(D_{n,\mathbb{C}}(\lambda)) = d_1(\Gamma_{n,\mathbb{C}}(\lambda), \Gamma_{\mathbb{C}}(\lambda))$. Write $\text{Leb}(D_{n,\mathbb{C}}(\lambda))$ as a sum of two terms:

$$\begin{aligned} \text{Leb}(D_{n,\mathbb{C}}(\lambda)) &= \text{Leb}(D_{n,\mathbb{C}}(\lambda) \cap \{x: |f(x) - \lambda| < \eta\}) + \text{Leb}(D_{n,\mathbb{C}}(\lambda) \cap \{x: |f(x) - \lambda| \geq \eta\}). \end{aligned}$$

The first term on the right-hand side is dominated by $F\{x: |f(x) - \lambda| < \eta\}$. As for the second term we have

$$H_{\lambda}(\Gamma(\lambda)) - H_{\lambda}(\Gamma_{n,\mathbb{C}}(\lambda)) = \int_{D_{n,\mathbb{C}}(\lambda)} |f(x) - \lambda| dx$$

$$\geq \eta \text{Leb}(D_{n,\mathbb{C}}(\lambda) \cap \{x: |f(x) - \lambda| \geq \eta \}).$$

It remains to show that

$$H_\lambda(\Gamma(\lambda)) - H_\lambda(\Gamma_{n,\mathbb{C}}(\lambda)) \leq (F_n - F)(\Gamma_{n,\mathbb{C}}(\lambda)) - (F_n - F)(\Gamma(\lambda)).$$

But this easily follows from $0 \leq H_{n,\lambda}(\Gamma_{n,\mathbb{C}}(\lambda)) - H_{n,\lambda}(\Gamma(\lambda))$ by using the fact that $H_{n,\lambda} = H_\lambda + F_n - F$.

□

Proof of Lemma 5.1: For every given set $\Lambda \subset [0, \infty)$ it follows immediately from the definition of $f(x, \Lambda)$ that $f(x) = f(x, \Lambda) + f(x, \Lambda^c)$. The analogous decomposition holds for $f_{n,\mathbb{C}}$. Hence,

$$\begin{aligned} d_I(f_{n,\mathbb{C}}, f) &\leq \int |f_{n,\mathbb{C}}(x, \Lambda) - f(x, \Lambda)| dx + \int f_{n,\mathbb{C}}(x, \Lambda^c) dx + \int f(x, \Lambda^c) dx \\ (5.9) \quad &= d_I(f_{n,\mathbb{C}}(\cdot, \Lambda), f(\cdot, \Lambda)) + \int f_{n,\mathbb{C}}(x, \Lambda^c) dx + \int f(x, \Lambda^c) dx. \end{aligned}$$

Now consider the second integral in the last line. Since $\int f_{n,\mathbb{C}}(x) dx = 1$ (Proposition 2.3) we have:

$$\begin{aligned} \int f_{n,\mathbb{C}}(x, \Lambda^c) dx &= 1 - \int f_{n,\mathbb{C}}(x, \Lambda) dx \\ &= \int f(x) dx - \int f_{n,\mathbb{C}}(x, \Lambda) dx \\ &= \int f(x, \Lambda) - f_{n,\mathbb{C}}(x, \Lambda) dx + \int f(x, \Lambda^c) dx \\ &\leq d_I(f_{n,\mathbb{C}}(\cdot, \Lambda), f(\cdot, \Lambda)) + \int f(x, \Lambda^c) dx. \end{aligned}$$

Together with (5.9) the assertion follows.

□

Proof of Proposition 5.2: The idea of the proof is the same as for the proof of Theorems 3.6 and 3.7 in Polonik [11]. However, since here we have to use the function-indexed empirical process (and not only the set-indexed empirical process as in Polonik [11]) we need several new arguments and therefore we give the complete proof.

To shorten the notation let $g_n(\lambda) = (|\varphi'(\xi_{\lambda,n})| \vee 1)^{-1}$, $\lambda \in (a,b)$. Choose $\eta = \alpha_n$ in (5.3). Then, by definition of $\xi_{\lambda,n}$, multiplication of (5.3) by $g_n(\lambda)$ leads to

$$g_n(\lambda) d_I(\Gamma_{n,\mathbb{C}}(\lambda), \Gamma(\lambda)) \leq 2 \alpha_n + \alpha_n^{-1} g_n(\lambda) [(F_n - F)(\Gamma_{n,\mathbb{C}}(\lambda)) - (F_n - F)(\Gamma(\lambda))]$$

Hence, we obtain

$$\begin{aligned} d_I(g_n(\lambda) \Gamma_{n,\mathbb{C}}(\lambda), g_n(\lambda) \Gamma(\lambda)) & \\ & \leq 2 \alpha_n + \alpha_n^{-1} [(F_n - F)(g_n(\lambda) \Gamma_{n,\mathbb{C}}(\lambda)) - (F_n - F)(g_n(\lambda) \Gamma(\lambda))], \\ (5.10) \quad & = 2 \alpha_n + \alpha_n^{-1} (F_n - F) [g_n(\lambda) (\Gamma_{n,\mathbb{C}}(\lambda) \setminus \Gamma(\lambda) - \Gamma_{n,\mathbb{C}}(\lambda) \setminus \Gamma(\lambda))] \end{aligned}$$

where we identify sets with their indicator functions and for any measure F and any integrable function g we denote $F(g) = \int g dF$. Let $G_{\mathbb{C}} = \{ r(C \setminus D - D \setminus C), r \leq 1, C, D \in \mathbb{C} \}$ and for a sequence $\{\delta_n\}$ of positive real numbers define

$$A_n = \{ \sup_{\lambda \in (a,b)} d_I(g_n(\lambda) \Gamma_{n,\mathbb{C}}(\lambda), g_n(\lambda) \Gamma(\lambda)) > 3 \delta_n \} \text{ and}$$

$$\begin{aligned} B_n = \{ \exists g \in G_{\mathbb{C}} \text{ such that } \|g\|_{1,Leb} > 3 \delta_n \text{ and} \\ \|g\|_{1,Leb} \leq 2 \alpha_n + 2 \alpha_n^{-1} |(F_n - F)(g)| \} \end{aligned}$$

where for any measure G on \mathbf{R}^d , $\|\cdot\|_{1,G}$ denotes the L_1 -norm with respect to G . Then, since for $C, D \in \mathbb{C}$ and $0 < r < 1$ we have $d_1(rC, rD) = \|r(C \setminus D - D \setminus C)\|_{1,Leb}$ it follows from (5.10) that $A_n \subset B_n$. Hence, in order to prove Proposition 5.2 we have to show that there exists a constant $C > 0$ such that for $\delta_n = C \alpha_n / 3$ we have $P^*(B_n) \rightarrow 0$ as $n \rightarrow \infty$. Note that with this choice of δ_n we have $B_n \subset C_n$, where

$$\begin{aligned} C_n = \{ \exists g \in G_{\mathbb{C}} \text{ such that } \|g\|_{1,Leb} > C \alpha_n \text{ and} \\ 1 \leq 2/C + 2 \alpha_n^{-1} |(F_n - F)(g)| / \|g\|_{1,Leb} \} \end{aligned}$$

We shall show that $P^*(C_n) \rightarrow 0$ as $n \rightarrow \infty$. In order to prove this it suffices to show that there exists a constant $C > 4$ such that

$$P^*(\sup_{g \in G_{\mathbb{C}}: \|g\|_{1,Leb} > 3C\alpha_n} |(F_n - F)(g)| / \|g\|_{1,Leb} > \alpha_n/4) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

In the following the supremum is always extended over $g \in G_{\mathbb{C}}$ which satisfy a certain condition. To shorten the notation we omit “ $g \in G_{\mathbb{C}}$ “. We have

$$\begin{aligned} & P^*(\sup_{\|g\|_{1,Leb} > C\alpha_n} |(F_n - F)(g)| / \|g\|_{1,Leb} > \alpha_n/2) \\ & \leq \sum_{j=0}^{\infty} P^*(\sup_{\|g\|_{1,Leb} \leq C2^j\alpha_n} |(F_n - F)(g)| > C2^{j-1}\alpha_n^2) \\ & \leq \sum_{j=0}^{\infty} P^*(\sup_{\|g\|_{1,Leb} \leq C2^j\alpha_n} |v_n(g)| > C2^{j-1}n^{1/2}\alpha_n^2) \\ & = \sum_{j=0}^{\infty} P_{n,j} \end{aligned}$$

where $v_n = n^{1/2}(F_n - F)$. In order to prove that the last sum converges to zero as n tends to infinity, we shall use results of Alexander [1]. For the case that \mathbb{C} is a VC-classes his results can be used directly. However, for the case that \mathbb{C} satisfies (3.3) we need to extend his results from the set-indexed to the function-indexed empirical process. We formulate this extension (Theorem 5.3, below) for a special case such that we can use it directly. For a class G of functions with $G \subset L_p(F)$ let

$$N_{p,B}(\varepsilon, G, F) := \inf \left\{ m \in \mathbb{N} : \exists g_1, \dots, g_m \text{ measurable, such that for every } g \in G \right. \\ \left. \text{there exist } i, j \in \{1, \dots, m\} \text{ with } g_i \leq g \leq g_j \text{ and } \|g_j - g_i\|_{p,F} < \varepsilon \right\},$$

where $\|\cdot\|_{p,F}$ denotes the L_p -norm with respect to F . Then $\log N_{p,B}(\varepsilon, G, F)$ is called metric entropy with bracketing of G in $L_p(F)$.

Theorem 5.3 (Alexander): *Let G be a class of functions with $0 \leq g \leq 1$. Let $L(x) = \log(x \vee e)$, $n \geq 1$, $\alpha \geq \sup_{g \in G} \text{var } v_n(g)$, with $v_n = n^{1/2}(F_n - F)$ and define $\Psi(L, n, \alpha) = L^2 / 2\alpha(1 + L/3n^{1/2}\alpha)$*

Part I: Suppose that G is a n -deviation measurable class of functions such that the graph region class $\{(x,t), g(x) \geq t \geq 0\}, g \in G$ is a (v,k) -constructible VC-class. There exist constants $K_i = K_i(v,k), i = 1,2$, such that if

$$L \geq \alpha^{1/2}$$

and

$$L > K_1 n^{-1/2} L(n)$$

and

$$L > K_2 (\alpha L(1/\alpha))^{1/2},$$

then

$$P^*(\sup_{g \in G} |v_n(g)| > L) \leq 5 \exp\{-1/2 \Psi(L,n,\alpha)\}.$$

Part II: Suppose that

$$\log N_{L,B}(\varepsilon, G, F) \leq A \varepsilon^{-r} \quad \forall \varepsilon > 0.$$

Then there exist constants $K_i = K_i(r,A), i = 1,2,3$, such that if

$$L \geq \begin{cases} K_1 \alpha^{(1-r)/2} & \text{if } r < 1 \\ K_2 L(n) & \text{if } r = 1 \\ K_3 n^{(r-1)/2(r+1)} & \forall r \end{cases}$$

then

$$P^*(\sup_{g \in G} |v_n(g)| > L) \leq 5 \exp\{-1/2 \Psi(L,n,\alpha)\}.$$

Part I of Theorem 5.3 directly follows from Theorem 2.8 of Alexander [1]. The proof of part II essentially is the same as the proof of Corollary 2.4 of Alexander [1]. The changes that have to be made will be outlined below after the proof of Proposition 5.2.

Remember that $M = \sup \{f(x)\} < \infty$. Let $C > 0$ be a constant. It will turn out that C can be chosen to be bigger than 2 as required. Define

$$G_{C,n,j} = \{g \in G_C: \|g\|_{L,F} < M C 2^j \alpha_n\},$$

then we have $p_{n,j} \leq P^*(\sup_{F_{\dots}} |v_n(f)| > C 2^{j-1} n^{1/2} \alpha_n^2)$.

Note that for any $g \in G_C$ we have $\text{var}(v_n(g)) \leq \|g\|_{2,F}^2 \leq \|g\|_{1,F} \leq M \|g\|_{1,\text{Leb}}$, so that

$$\sup_{g \in G_{C,n,j}} \text{var}(v_n(g)) \leq M C 2^j \alpha_n.$$

We now want to apply Theorem 5.3 for fixed n and j with $\alpha = M C 2^j \alpha_n$ and $L = C 2^{j-1} n^{1/2} \alpha_n^2$. First note, that if $\log N_1(\epsilon, \mathbb{C}, F) = O(\epsilon^{-r})$, then $\log N_{1,B}(\epsilon, G_C, F)$ is of the same order. This is easy to verify. Now one has to check that the conditions of Theorem 5.3 are satisfied with this choice of α and L . This is an easy calculation, which in addition shows, that the conditions are satisfied for all $C > C_0$, C_0 large enough, independent of n and j . Hence, as required, C can be chosen to be bigger than 2. It follows that

$$\begin{aligned} \sum_{j=1}^{\infty} p_{n,j} &\leq 5 \sum_{j=1}^{\infty} \exp\{-1/2 \Psi(C 2^{j-1} n^{1/2} \alpha_n^2, n, M C 2^j \alpha_n)\} \\ &= 5 \sum_{j=1}^{\infty} \exp\{-(C 2^j n \alpha_n^3) / (8M (1 + \alpha_n/6M))\}. \end{aligned}$$

Since $\alpha_n \rightarrow 0$ as $n \rightarrow \infty$ and $n \alpha_n^3 \geq \log n$ it follows that $\sum_{j=1}^{\infty} p_{n,j} \rightarrow 0$ as $n \rightarrow \infty$. □

Remarks on the proof of Theorem 5.3, Part II: The proof of Alexander [1], Corollary 2.4, goes through almost word by word, if one replaces δ_j by δ_j^2 and uses L_1 -bracketing functions instead of L_2 -bracketing functions. (In constructing the δ_j one has to use Lemma 3.1 with $H(x) = \log N_{1,B}(x^2, G, F)$ instead of $H(x) = \log N_{2,B}(x, G, F)$).

Alexander formulated this result only for the case that G is a class of sets, where he gave the condition on the metric entropy in terms of the L_2 -bracketing numbers. However, for classes of sets one has $\|g\|_{2,F}^2 = \|g\|_{1,F}$, so that $N_{2,B}(\epsilon, G, F) = N_{1,B}(\epsilon^2, G, F)$. Hence, for classes of sets it does in principle no matter if the conditions are formulated in L_1 -or in L_2 -bracketing numbers. However, for function classes G with $0 \leq g \leq 1$ for all $g \in G$ one has $\|g\|_{2,F}^2 \leq \|g\|_{1,F}$, so that, heuristically speaking, in order to control the L_1 - as well as the L_2 -norm of functions in G , one has to give conditions in terms of the L_1 -norm. □

Acknowledgement: This work is a revised version of a part of my thesis, which was written under the supervision of Prof. D.W. Müller at the Universität Heidelberg. I would like to thank Prof. Müller for his interest and support during the time of writing my thesis. Furthermore I thank the statistic group of the Universität Heidelberg, in particular W. Ehm, for valuable discussions, hints and remarks concerning the subject.

References:

- [1] Alexander, K.S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.* **12** 1041-1067.
- [2] Dudley, R.M. (1984). A course on empirical processes, *École d'Été de Probabilités de Saint Flour XII-1982, Lecture Notes in Math.* **1097** 1-142. New York: Springer.
- [3] Eddy, W.F. Hartigan, J.A. (1977). Uniform convergence of the empirical distribution function over convex sets. *Ann. Statist.* **5** 370-374.
- [4] Hartigan, J.A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.* **82** 267-270.
- [5] Grenander, U. (1956). On the theory of mortality measurement, Part II. *Skand. Akt.* **39** 125-153.
- [6] Groeneboom, P. (1985). Estimating a monotone density. in *Proceedings of the Berkeley Conference in honor of Jerzy Neymann and Jack Kiefer*, Vol. II, LeCam, L. Olshen, R. (eds.), Monterey: Wadsworth.
- [7] Müller, D.W. and Sawitzki, G. (1987). Using excess mass estimates to investigate the modality of a distribution. Preprint Nr.398, SFB 123, Universität Heidelberg.

- [8] Müller, D.W. and Sawitzki, G. (1991). Excess mass estimates and tests of multimodality. *J. Amer. Statist. Assoc.* **86** 738-746.
- [9] Nolan, D. (1991). The excess-mass ellipsoid. *J. Multivariate Anal.* **39** 348 - 371.
- [10] Polonik, W. (1992). The excess mass approach to cluster analysis and related estimation procedures. Dissertation, Universität Heidelberg.
- [11] Polonik, W. (1993). Measuring mass concentrations and estimating density contour clusters - an excess mass approach. Beiträge zur Statistik Nr. 7, Institut für Angewandte Mathematik, Universität Heidelberg.
- [12] Sager, T.W. (1979). An iterative method for estimating a multivariate mode and isopleth. *J. Amer. Statist. Assoc.* **74** 329-339.