Available online at www.sciencedirect.com
ScienceDirect

Journal of Multivariate Analysis 98 (2007) 1751 – 1764

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

Entropy estimate for high-dimensional monotonic functions

Fuchang Gao[a,*,1], Jon A. Wellner[b,2]

[a]*Department of Mathematics, University of Idaho, P.O. Box 441103, Moscow, ID 83844-1103, USA*
[b]*Department of Statistics, University of Washington, P.O. Box 354322. Seattle, WA 98195-4322, USA*

Received 29 December 2005
Available online 30 October 2006

Abstract

We establish upper and lower bounds for the metric entropy and bracketing entropy of the class of d-dimensional bounded monotonic functions under L^p norms. It is interesting to see that both the metric entropy and bracketing entropy have different behaviors for $p < d/(d-1)$ and $p > d/(d-1)$. We apply the new bounds for bracketing entropy to establish a global rate of convergence of the MLE of a d-dimensional monotone density.

AMS 1991 subject classification: 41A25; 62G05

Keywords: Block decreasing density; Metric entropy; Bracketing entropy; Maximum likelihood estimator

1. Introduction

Shape constrained functions appear very commonly in non-parametric estimation in statistics via renewal theory and mixing of uniform distributions. A class of multivariate functions of interest in applications is the class of “block-decreasing” densities; see e.g., [15,16,1]. It consists of bounded densities on $\mathbb{R}^d$ that are decreasing in each variable. We denote by $\mathcal{F}_d$ the collection of non-negative functions on $[0, 1]^d$ which are bounded by 1, and monotonic in each variable, that is, monotonic along any line that is parallel to an axis. As is well known, the rate of convergence of non-parametric estimators such as the maximum likelihood estimator (MLE) is determined by

* Corresponding author. Fax: +1 208 885 5843.
E-mail address: fuchang@uidaho.edu (F. Gao).

[1] Supported in part by NSF Grant DMS-0405855.
[2] Supported in part by NSF Grant DMS-0503822.

0047-259X/$ - see front matter
doi:10.1016/j.jmva.2006.09.003

the metric entropy and bracketing entropy bounds for an appropriate related class of functions; see the definitions below.

In this paper, we provide upper and lower bounds for the metric entropy $\log N(\varepsilon, \mathcal{F}_d, \|\cdot\|_p)$ and the bracketing entropy $\log N_{[\,]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_p)$ for $1 \leqslant p < \infty$, where $\|\cdot\|_p$ is the L^p norm under Lebesgue measure, $N(\varepsilon, \mathcal{F}_d, \|\cdot\|_p)$ and $N_{[\,]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_p)$ are defined as follows:

$$N(\varepsilon, \mathcal{F}_d, \|\cdot\|_p) := \min\left\{ m : \exists f_1, f_2, \ldots, f_m \text{ s.t. } \mathcal{F}_d \subset \bigcup_{k=1}^{m} B_p(f_k, \varepsilon) \right\},$$

where $B_p(f_k, \varepsilon) = \{f \in \mathcal{F}_d : \|f - f_k\|_p \leqslant \varepsilon\}$, and

$$N_{[\,]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_p) := \min\left\{ m : \exists \underline{f}_1, \overline{f}_1, \ldots, \underline{f}_m, \overline{f}_m \in L_d^0 \text{ s.t. } \|\overline{f}_k - \underline{f}_k\|_p \leqslant \varepsilon, \mathcal{F}_d \subset \bigcup_{k=1}^{m} [\underline{f}_k, \overline{f}_k] \right\},$$

where L_d^0 is the set of all measurable functions on $[0, 1]^d$, and

$$[\underline{f}_k, \overline{f}_k] = \left\{ g \in \mathcal{F}_d : \underline{f}_k \leqslant g \leqslant \overline{f}_k \right\}.$$

Note that in the definition of $N(\varepsilon, \mathcal{F}_d, \|\cdot\|_p)$ we do not require the centers of the balls $f_1, \ldots, f_n$ to belong to $\mathcal{F}_d$. Likewise, in the definition of $N_{[\,]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_p)$, the functions $\underline{f}_k$ and $\overline{f}_k$ do not have to belong to $\mathcal{F}_d$.

Our main result is the following:

Theorem 1.1. *For $1 \leqslant p < \infty$ and $d \geqslant 2$, there exist constants c_1 and c_2 depending only on p and d, such that if $(d-1)p \neq d$, then*

$$c_1 \varepsilon^{-\alpha} \leqslant \log N(\varepsilon, \mathcal{F}_d, \|\cdot\|_p) \leqslant \log N_{[\,]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_p) \leqslant c_2 \varepsilon^{-\alpha},$$

where $\alpha = \max\{d, (d-1)p\}$. If $(d-1)p = d$, then

$$c_1 \varepsilon^{-d} \leqslant \log N(\varepsilon, \mathcal{F}_d, \|\cdot\|_p) \leqslant \log N_{[\,]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_p) \leqslant c_2 \varepsilon^{-d} (\log 1/\varepsilon)^d. \tag{1}$$

The new bracketing entropy bounds have implications for the rate of convergence of the MLE of a "block decreasing" density as will be shown in Section 5.

Remark 1.2. We believe that in the critical case $(d-1)p = d$, the logarithmic factor in the upper bound in (1) is not needed, and prove in Theorem 4.1 that this is indeed so for regular metric entropy under the L^p norm, provided $(d, p) \neq (2, 2)$. The challenge in the case $p = d = 2$ is somewhat surprising. We conjecture that $c_1 \varepsilon^{-2} \leqslant \log N(\varepsilon, \mathcal{F}_2, \|\cdot\|_2) \leqslant c_2 \varepsilon^{-2}$ for some positive constants c_1 and c_2.

Now, let us point out some connections with known results:

Theorem 1.1 is related to known results as follows:

(i) When $d = 1$, $\mathcal{F}_d$ is just the class of probability distribution functions, and the entropies are known to be of the order ε^{-1}; see e.g. [18, Theorem 2.75, p. 159]. So, in some sense, the results in this paper generalize the known results for $d = 1$.

(ii) The result for the case $p = 1$ is also known. Indeed, it is easy to see that if f is a non-negative block-decreasing function on $\mathbb{R}^d$, then the set $B = \{(x_1, \ldots, x_d, y) \in \mathbb{R}^{d+1} : y \leqslant f(x_1, \ldots, x_d)\}$ satisfies the following condition: for all $x = (x_1, \ldots, x_{d+1}) \in B$ and $z = (z_1, \ldots, z_{d+1})$ with $z_1 \leqslant x_1, \ldots, z_{d+1} \leqslant x_{d+1}$, we have $z \in B$. Such sets are called lower layers in [11,12]. Similarly, the sets $\{x \in [0, 1]^d : f(x) \geqslant t\}$, $t > 0$, are lower layers in $[0, 1]^d$. On the other hand, if B is a lower layer set with $B \cap [0, 1]^{d+1} \neq \emptyset$, then the function defined by

$$f(x_1, \ldots, x_d) = \sup\{y : (x_1, \ldots, x_d, y) \in B \cap [0, 1]^{d+1}\}$$

is block-decreasing and bounded by 1. In the case $p = 1$, metric entropy bounds for the collection of lower layers were studied in [12, Theorem 8.3.2, p. 266]. Thus, Theorem 1.1 can also be viewed as the L^p version of entropy bound of lower layers. These connections were exploited by Polonik [15,16] as will be discussed further in Section 5.

(iii) The result for the upper bound of the regular entropy estimate in the case $p < d/(d-1)$ can be obtained by applying a general result for Besov spaces. In fact, for any $f \in \mathcal{F}_d$, because f is bounded and monotonic along each line that is parallel to an axis, it is of bounded variation in each variable. Therefore, by [19, Theorem 5.3.5] f is of bounded variation. From this, we immediately obtain that $\mathcal{F}_d \subset B^1_{1,\infty}([0, 1]^d)$ (cf. [13, Section 2.2.2 (12)]). Now, a general result on entropy in Besov spaces (cf. [13, Theorem 3.5], or [7, Corollary 1]) gives an upper bound for the regular entropy in the case $p < d/(d-1)$. This connection also partially explains why the entropy behaves differently in the case $p < d/(d-1)$ and the case $p > d/(d-1)$.

(iv) It should also be noted that when $d > 1$, $\mathcal{F}_d$ is much larger than the class $\mathcal{D}_d$ of d-dimensional probability distributions. Indeed, a block increasing function is a function with non-negative and bounded first order partial differences; while a d-dimensional distribution is a function of non-negative and bounded mixed partial differences up to order d. For example, the block increasing function $f(x, y) = x + y - xy$ on $[0, 1]^2$ is not a probability distribution because its mixed partial difference is negative. The boundedness and non-negativity of the mixed partial difference have a huge impact on the entropy bounds even when $d = 2$. Indeed, our result in the case $p = d = 2$ says

$$c_1\varepsilon^{-2} \leqslant \log N(\varepsilon, \mathcal{F}_2, \|\cdot\|_2) \leqslant c_2\varepsilon^{-2}[\log(1/\varepsilon)]^2$$

for some positive constants c_1 and c_2; while Blei et al. [8] recently proved that

$$c_1\varepsilon^{-1}[\log(1/\varepsilon)]^{3/2} \leqslant \log N(\varepsilon, \mathcal{D}_2, \|\cdot\|_2) \leqslant c_2\varepsilon^{-1}[\log(1/\varepsilon)]^{3/2}.$$

As d increases, the difference becomes even bigger.

The results of this paper have a number of statistical implications. This is clarified by the following corollary giving upper bounds on bracketing entropies measured with respect to the $\|\cdot\|_{Q,p}$ norm where $\|f\|_{Q,p} \equiv \{Q|f|^p\}^{1/p}$.

Corollary 1.3. *Suppose Q is a probability measure on $[0, 1]^d$ with bounded Lebesgue density q; thus $\|q\|_\infty = \sup_{x\in I^d} q(x) < \infty$. Then if $(d-1)p \neq d$ there is a constant $C_2 = C_2(p, \|q\|_\infty)$ such that*

$$\log N_{[\,]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_{Q,p}) \leqslant C_2\varepsilon^{-\alpha},$$

where $\alpha = \max\{d, (d-1)p\}$. *If* $(d-1)p = d$, *then*

$$\log N_{[\,]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_{Q,p}) \leqslant C_2 \varepsilon^{-d} (\log 1/\varepsilon)^d$$

for a (different) constant $C_2 = C_2(p, \|q\|_\infty)$.

By the bracketing Glivenko–Cantelli theorem (see e.g. [18, Theorem 2.4.1, p. 122] or [12, Theorem 7.1.5, p. 235]), Corollary 1.3 with $p = 1$ implies that $\mathcal{F}_d$ is a P_0-Glivenko–Cantelli class for any probability measure P_0 on $[0,1]^d$ with bounded Lebesgue density p_0. The bracketing entropy estimate (and its proof) gives the lower bound estimate of the shattering dimension of $\mathcal{F}_d$ [17], that is $v(\mathcal{F}_d, t) \geqslant ct^{-(2d-2)}$, and lower bound estimate of Kolchinskii–Pollard entropy $D(\mathcal{F}_d, t)$, that is $D(\mathcal{F}_d, t) \geqslant ct^{-(2d-2)}$. When $d > 2$, $v(\mathcal{F}_d, t) \neq O(t^{-2})$ and $D(\mathcal{F}_d, t) \neq O(t^{-2})$. So by [17, Theorem 1.1] or the Sudkov–Chevet Theorem ([12, Theorem 2.3.5, p. 33]) we immediately conclude that $\mathcal{F}_d$ is not a P-Donsker class when $d > 2$ and P is the uniform (i.e. Lebesgue) measure on $[0,1]^d$. In the case $d = 2$, it follows from the results of [12, Section 12.4, pp. 373–388], that $\mathcal{F}_2$ is also not P-Donsker for P the uniform distribution: Letting $\mathcal{F}_2$ denote the collection of non-negative block decreasing functions bounded by 1 on $[0,1]^d$, we have $\mathcal{F}_2 \supset \{1_C : C \in \mathcal{LL}_{2,1}\}$ using the notation of [12, Section 8.3]. Thus, with $\mathbb{G}_n \equiv \sqrt{n}(\mathbb{P}_n - P)$ and $\|\mathbb{G}_n\|_{\mathcal{F}} \equiv \sup_{f \in \mathcal{F}} |\mathbb{G}_n(f)|$, we have

$$\|\mathbb{G}_n\|_{\mathcal{F}_2} \geqslant \|\mathbb{G}_n\|_{\mathcal{LL}_{2,1}},$$

and hence by [12, Theorem 12.4.2, p. 375], for the uniform distribution P on $[0,1]^2$, for every $\varepsilon > 0$ there is a $\delta > 0$ such that

$$Pr(\|\mathbb{G}_n\|_{\mathcal{F}_2} \geqslant \delta(\log n)^{3/4}) \geqslant Pr(\|\mathbb{G}_n\|_{\mathcal{LL}_{2,1}} \geqslant \delta(\log n)^{3/4}) \geqslant 1 - \varepsilon$$

for all n sufficiently large. This contradicts $\mathcal{F}_2$ being P-Donsker for the uniform distribution P. (See [12, p. 375], for further discussion of the optimality of the $(\log n)^{3/4}$ term in this argument.)

The remainder of the paper is organized as follows. First, we prove the lower bound for regular entropy by constructing a well-separated set using a combinatorial argument. Next, we obtain the upper bound for bracketing entropy using a constructive proof, revealing the difference of entropy growth between the cases $p < d/(d-1)$ and $p > d/(d-1)$. Then we turn to the critical case $p = d/(d-1)$, and use the result for the case $p = 1$ and the metric entropy estimate of convex hulls to remove the extra logarithmic factor in the upper bound for the regular entropy. Finally, we apply the bracketing entropy estimate to establish a global rate of convergence of the MLE of a d-dimensional "block-decreasing" density.

2. Lower bound

In this section, we obtain the lower bound estimate, namely

Proposition 2.1. *For* $p \geqslant 1$, *there exists a constant* $c_1 > 0$ *such that*

$$\log N(\varepsilon, \mathcal{F}_d, \|\cdot\|_p) \geqslant c_1 \varepsilon^{-\alpha},$$

where $\alpha = \max\{d, (d-1)p\}$.

Proof. For convenience, we assume $\varepsilon = 2^{-n}$ for some positive integer n. We divide $[0,1]^d$ into ε^{-d} small cubes of side-length ε. Define g on $[0,1]^d$, such that on each open cube

$\prod_{i=1}^{d}(k_i\varepsilon, k_i\varepsilon+\varepsilon), 0\leqslant k_i < 2^n, 1\leqslant i\leqslant d,$

$$g(x) = \frac{(k_1+k_2+\cdots+k_d+1)\varepsilon}{3d} \pm \frac{\varepsilon}{6d}.$$

Clearly, there are $2^{\varepsilon^{-d}}$ different ways to define g, and each can be extended to a function in $\mathcal{F}_d$. Let $\mathcal{G}_d$ be the collection of these extended functions.

For each $g \in \mathcal{G}_d$ define

$$B(g) = \{h \in \mathcal{G}_d\colon \text{ there are at most } 2^{-4}\varepsilon^{-d} \text{ open cubes on which } g \neq h\}.$$

Since $\binom{m}{l}\leqslant(me/l)^l$ and $(16e)^{1/16}\leqslant 2^{1/2}$, it is easy to check that $B(g)$ contains no more than $\binom{\varepsilon^{-d}}{2^{-4}\varepsilon^{-d}}\leqslant 2^{\varepsilon^{-d}/2}$ elements. Thus, we can find $N = 2^{\varepsilon^{-d}/2}$ functions $g_1, g_2, \ldots, g_N$, such that if $i \neq j$, then $B(g_i)$ and $B(g_j)$ are disjoint. Clearly

$$\|g_i - g_j\|_1 \geqslant \frac{\varepsilon}{3d}\cdot\frac{1}{2^4} = \frac{\varepsilon}{48d}.$$

Hence, $N((48d)^{-1}\varepsilon, \mathcal{F}_d, \|\cdot\|_1)\geqslant 2^{\varepsilon^{-d}/2}$, which implies

$$N(\varepsilon, \mathcal{F}_d, \|\cdot\|_p,)\geqslant N(\varepsilon, \mathcal{F}_d, \|\cdot\|_1,)\geqslant e^{c_1\varepsilon^{-d}}$$

for some constant $c_1 > 0$ and all $p\geqslant 1$.

When $p > d/(d-1)$, this lower bound is not sharp. In order to improve it, we will construct a different well-separated subset. We define $q(x)$ on $[0,1]^d$ as follows: on each open cube $\prod_{i=1}^{d}(k_i\varepsilon, k_i\varepsilon+\varepsilon)$ that satisfies $k_1+k_2+\cdots+k_d = \varepsilon^{-1}, k_1, k_2, \ldots, k_d\geqslant 0$, we define $q(x) = \frac{1}{2}\pm\frac{1}{2}$. Clearly, $q(x)$ can be extended to a function in $\mathcal{F}_d$. Now, because there are $c\varepsilon^{1-d}$ qualified cubes, where c is a constant depending only on d, there are $2^{c\varepsilon^{1-d}}$ different functions $q(x)$. The same combinatorial argument as the one given above shows that there are at least $m = 2^{c\varepsilon^{1-d}/2}$ functions $q_1, q_2, \ldots, q_m$, such that $|q_i - q_j| = 1$ on at least $c\varepsilon^{1-d}/2^4$ cubes, $i \neq j$. Thus,

$$\|q_i - q_j\|_p \geqslant \left(\frac{c\varepsilon}{2^4}\right)^{1/p}.$$

This implies that

$$N((c2^{-4}\varepsilon)^{1/p}, \mathcal{F}_d, \|\cdot\|_p)\geqslant 2^{c\varepsilon^{1-d}/2},$$

which further implies

$$N(\varepsilon, \mathcal{F}_d, \|\cdot\|_p)\geqslant e^{c_1\varepsilon^{-(d-1)p}},$$

for some constant $c_1 > 0$ when $p > d/(d-1)$. □

3. Upper bound

In this section, we obtain an upper bound through a constructive proof. We will prove

Proposition 3.1. *For $p\geqslant 1$, $p \neq d/(d-1)$, there exists a constant $c_2 > 0$ such that*

$$\log N_{[\,]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_p)\leqslant c_2\varepsilon^{-\alpha},$$

where $\alpha = \max\{d, (d-1)p\}$. *For* $p = d/(d-1)$, *there exists a constant* $c_2 > 0$ *such that*

$$\log N_{[\,]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_p) \leqslant c_2 \varepsilon^{-d} (\log 1/\varepsilon)^d.$$

3.1. Construction

For convenience, we introduce the notion

$$\omega(f, I) = \sup\{f(t) : t \in I\} - \inf\{f(t) : t \in I\},$$

where I is any subset of $[0, 1]^d$.

If $p = 1$, we choose $K = 2^d$; otherwise, we choose $K = 2^\beta$, where $\beta = \frac{1}{2}[d - 1 + 1/(p-1)]$. For any given $\varepsilon = 2^{-n}$, $n \in \mathbb{N}$, let l be the integer satisfying $K^{-l} \leqslant \varepsilon < K^{-l+1}$.

For each $f \in \mathcal{F}_d$, we construct $\underline{f}$ and $\overline{f}$ as follows. First, we partition $[0, 1)^d$ into ε^{-d} cubes of side-length ε. (All the cubes are of the form $\prod_{i=1}^d [a_i, b_i)$.) A cube I_0 of side-length ε is selected if $\omega(f, I_0) \leqslant K\varepsilon$. For each cube that is not selected, we partition it into 2^d cubes of equal size. In general, suppose we have a cube I_i of side-length $2^{-i}\varepsilon$. If $\omega(f, I_i) \leqslant K^{i+1}\varepsilon$, we select the cube; otherwise, we partition the cube into 2^d smaller cubes. This process continues until $i = l$. In this case, we always select the cube. Clearly, each point in $[0, 1)^d$ uniquely belongs to one of the selected cubes.

On each selected cube I of side-length $2^{-i}\varepsilon$, $0 \leqslant i < l$, we define

$$\underline{f} = K^{i+1}\varepsilon \left\lfloor \frac{\inf_{x \in I} f(x)}{K^{i+1}\varepsilon} \right\rfloor, \quad \overline{f} = K^{i+1}\varepsilon \left\lceil \frac{\sup_{x \in I} f(x)}{K^{i+1}\varepsilon} \right\rceil.$$

On each selected cube of side-length $2^{-l}\varepsilon$ and on $[0, 1]^d \setminus [0, 1)^d$, we define $\overline{f} = 1$ and $\underline{f} = 0$. Clearly, $\underline{f} \leqslant f \leqslant \overline{f}$.

Let $\overline{\mathcal{S}} = \{\overline{f} : f \in \mathcal{F}_d\}$, and $\underline{\mathcal{S}} = \{\underline{f} : f \in \mathcal{F}_d\}$. We will estimate $\|\overline{f} - \underline{f}\|_p$, and the cardinalities $|\underline{\mathcal{S}}|$ and $|\overline{\mathcal{S}}|$ of $\underline{\mathcal{S}}$ and $\overline{\mathcal{S}}$, respectively.

3.2. Bound for $\|\overline{f} - \underline{f}\|_p$

For each $i \in \mathbb{N}$, let U_i be the union of the selected cubes of side-length $2^{-i}\varepsilon$. We first bound the measure of U_i.

Let s_i be the number of cubes of side-length $2^{-i}\varepsilon$ that have been selected, and n_i be the number of cubes of side-length $2^{-i}\varepsilon$ that have not been selected. Clearly, by the construction of $\underline{f}$ and $\overline{f}$, we have $s_i + n_i = 2^d n_{i-1}$. In particular, $s_i \leqslant 2^d n_{i-1}$.

Now we try to estimate n_{i-1} for $i \geqslant 1$. If a cube $I = \prod_{j=1}^d [a_j, b_j)$ of side-length $2^{-i+1}\varepsilon$ is not selected, then $\omega(f, I) > K^i\varepsilon$. By the monotonicity of f along each variable, there exists $1 \leqslant j \leqslant d$, such that on the edge $\overline{A_{j-1}A_j}$, we have $\omega(f, \overline{A_{j-1}A_j}) > K^i\varepsilon/d$, where

$$A_j = (b_1, \ldots, b_j, a_{j+1}, \ldots, a_d).$$

Thus, for n_{i-1} cubes of side-length $2^{-i+1}\varepsilon$, there are n_{i-1} disjoint edges on which $\omega(f, \cdot) > K^i\varepsilon/d$. From these edges, there are at least $\lceil n_{i-1}/d \rceil$ edges that are parallel. Furthermore, from these parallel edges, there are at least $\lceil n_{i-1}(2^{-i+1}\varepsilon)^{d-1}/d \rceil$ disjoint edges that lie on the same line segment $[0, 1]$ that is parallel to one of the axes. Because f is monotonic along this line segment, and the value change is at most 1, we have

$$\lceil n_{i-1}(2^{-i+1}\varepsilon)^{d-1}/d\rceil \cdot \frac{K^i\varepsilon}{d} \leqslant 1.$$

Thus, $n_{i-1} \leqslant d^2 2^{(i-1)(d-1)} K^{-i}\varepsilon^{-d}$.

Therefore, for $1 \leqslant i \leqslant l$, the measure of U_i is bounded above by

$$\begin{aligned} s_i \cdot (2^{-i}\varepsilon)^d &\leqslant 2^d n_{i-1} \cdot (2^{-i}\varepsilon)^d \\ &\leqslant 2^d \cdot d^2 2^{(i-1)(d-1)} K^{-i}\varepsilon^{-d} \cdot (2^{-i}\varepsilon)^d \\ &= 2d^2(2K)^{-i}. \end{aligned}$$

For $i = 0$, the measure of U_0 is trivially bounded by 1.

Recall that for $0 \leqslant i < l$, $|\overline{f} - \underline{f}| \leqslant 2K^{i+1}\varepsilon$ on U_i. Also, on U_l, we have $|\overline{f} - \underline{f}| \leqslant 1$. Thus,

$$\begin{aligned} \|\overline{f} - \underline{f}\|_p^p &= \int_{U_0} |\overline{f} - \underline{f}|^p + \sum_{i=1}^{l-1} \int_{U_i} |\overline{f} - \underline{f}|^p + \int_{U_l} |\overline{f} - \underline{f}|^p \\ &\leqslant (2K\varepsilon)^p + \sum_{i=1}^{l-1} (2K^{i+1}\varepsilon)^p \cdot 2d^2(2K)^{-i} + 2d^2(2K)^{-l} \\ &\leqslant (2K\varepsilon)^p + 2^{p+1}K^p d^2 \sum_{i=1}^{l-1} \left(\frac{K^{p-1}}{2}\right)^i \varepsilon^p + 2d^2(2K)^{-l}. \end{aligned} \tag{2}$$

When $(d-1)p < d$, we have $d - 1 < \beta < \frac{1}{p-1}$. So, $K = 2^\beta < 2^{1/(p-1)}$. Thus, $K^{p-1}/2 < 1$, and $\frac{1}{2K} \leqslant K^{-p}$. Therefore

$$\begin{aligned} \|\overline{f} - \underline{f}\|_p^p &\leqslant (2K\varepsilon)^p + 2^{p+1}K^p d^2 \cdot \frac{K^{p-1}}{2 - K^{p-1}}\varepsilon^p + 2d^2 \cdot K^{-pl} \\ &\leqslant \left[(2K)^p + 2^{p+1}K^p d^2 \cdot \frac{K^{p-1}}{2 - K^{p-1}} + 2d^2\right]\varepsilon^p \\ &\leqslant c\varepsilon^p \end{aligned} \tag{3}$$

for some constant c depending only on p and d, where in the second inequality we used the fact that $K^{-l} \leqslant \varepsilon$.

When $(d-1)p > d$, we have $d - 1 > \beta > \frac{1}{p-1}$. So, $K = 2^\beta > 2^{1/(p-1)}$, that is $K^{p-1}/2 > 1$. Hence,

$$\begin{aligned} \|\overline{f} - \underline{f}\|_p^p &\leqslant (2K\varepsilon)^p + 2^{p+1}K^p d^2 \cdot \frac{(K^{p-1}/2)^l}{K^{p-1}/2 - 1}\varepsilon^p + 2d^2 \cdot (2K)^{-l} \\ &\leqslant (2K\varepsilon)^p + \frac{2^{p+1}K^p d^2}{K^{p-1}/2 - 1} \cdot K^{pl}\varepsilon^p \cdot (2K)^{-l} + 2d^2 \cdot (2K)^{-l} \\ &\leqslant (2K\varepsilon)^p + c(2K)^{-l} \\ &\leqslant (2K)^p\varepsilon^p + c\varepsilon^{1+1/\beta} \\ &\leqslant c'\varepsilon^{1+1/\beta}, \end{aligned}$$

for some constants $c, c' > 0$ depending only on p and d, where in the third and fourth inequalities we used the fact $1 \leqslant K^l \varepsilon < K$ and in last inequality we used the fact that $p > 1 + 1/\beta$.

When $(d-1)p = d$, we have $K^{p-1} = 2$. So, we obtain from (2) that

$$\begin{aligned}\|\overline{f} - \underline{f}\|_p^p &\leqslant (2K\varepsilon)^p + 2^{p+1}K^p d^2(l-1)\varepsilon^p + 2d^2(K^p)^{-l}\\ &\leqslant c\varepsilon^p \log 1/\varepsilon,\end{aligned}$$

for some constant $c > 0$ depending only on p and d, where in the last inequality we used the fact that $1 \leqslant K^l \varepsilon < K$.

Summarizing, we obtain that

$$\|\overline{f} - \underline{f}\|_p \leqslant \begin{cases} c\varepsilon & (d-1)p < d, \\ c\varepsilon(\log\, 1/\varepsilon)^{1/p} & (d-1)p = d, \\ c\varepsilon^{\frac{\beta+1}{p\beta}} & (d-1)p > d. \end{cases} \tag{4}$$

3.3. Bounds for $|\overline{\mathcal{S}}|$ and $|\underline{\mathcal{S}}|$

We derive the upper bound for $|\overline{\mathcal{S}}|$. The argument for bounding $|\underline{\mathcal{S}}|$ is almost identical.

Because all the selected cubes of side-length ε are chosen from $n_0 = \varepsilon^{-d}$ cubes, there are no more than $2^{\varepsilon^{-d}}$ different ways of selecting cubes of side-length ε. For $1 \leqslant i < l$, the selected cubes of side-length $2^{-i}\varepsilon$ are chosen from the n_{i-1} cubes of side-length $2^{-i+1}\varepsilon$ that were not selected in the previous step, there are no more than $2^{2^d n_{i-1}}$ different ways to select the cubes of side-length $2^{-i}\varepsilon$. Once the cubes are selected. For each $0 \leqslant i < l$, the s_i selected cubes of side-length $2^{-i}\varepsilon$ can be grouped into no more than $(2^i\varepsilon^{-1})^{d-1}$ rows. Suppose row-j contains r_j selected cubes. Because the values of $\overline{f}$ on these r_j cubes are in monotonic order, and are all chosen from $0, K^i\varepsilon, 2K^i\varepsilon, \ldots, mK^i\varepsilon$, where $m = \lfloor K^{-i}\varepsilon^{-1}\rfloor$, the number of different ways of assigning values of $\overline{f}$ on these r_j cubes is bounded by

$$\binom{r_j + \lfloor K^{-i}\varepsilon^{-1}\rfloor}{\lfloor K^{-i}\varepsilon^{-1} + 1} \leqslant \max\{\exp(cr_j), \exp(cK^{-i}\varepsilon^{-1})\} < \exp(cr_j)\cdot\exp(cK^{-i}\varepsilon^{-1}).$$

Thus, the number of different ways to assign the values of $\overline{f}$ on the s_i selected cubes of side-length $2^{-i}\varepsilon$ is bounded by

$$\begin{aligned}\prod_{j=1}^{(2^i\varepsilon^{-1})^{d-1}} (\exp(cr_j)\cdot\exp(cK^{-i}\varepsilon^{-1})) &\leqslant \exp(cs_i)\cdot\exp(c(2^{d-1}K^{-1})^i\varepsilon^{-d})\\ &\leqslant \exp(c'(2^{d-1}K^{-1})^i\varepsilon^{-d}),\end{aligned}$$

where in the inequality above, we used $s_i \leqslant 2^d n_{i-1}$, and the estimate $n_{i-1} \leqslant d^2 2^{(i-1)(d-1)}K^{-i}\varepsilon^{-d}$ obtained in Section 3.2.

Hence, the total number of realizations of $\overline{f}$ is bounded by

$$2^{\varepsilon^{-d}} e^{c'\varepsilon^{-d}} \prod_{i=1}^{l-1}\left[2^{2^d n_{i-1}}\cdot\exp\left(c'(2^{d-1}K^{-1})^i\varepsilon^{-d}\right)\right] \leqslant \exp\left(c''\sum_{i=0}^{l-1}(2^{d-1}K^{-1})^i\varepsilon^{-d}\right), \tag{5}$$

where in the last inequality we again used the estimate $n_{i-1} \leqslant d^2 2^{(i-1)(d-1)}K^{-i}\varepsilon^{-d}$.

When $(d-1)p > d$, $2^{d-1} > 2^{\beta} = K$, we can bound the right-hand side of (5) by

$$\exp\left(c'''[2^{d-1}/K]^l \varepsilon^{-d}\right) \leqslant \exp\left(c'''\varepsilon^{-(\beta+1)(d-1)/\beta}\right).$$

When $(d-1)p = d$, the upper bound of the right-hand side of (5) can be bounded by $\exp(c'''\varepsilon^{-d}\log 1/\varepsilon)$.

When $(d-1)p < d$, $2^{d-1}/K < 1$, and the upper bound of the right hand of (5) is bounded by $\exp(c'''\varepsilon^{-d})$.

Summarizing, we obtain

$$\log|\overline{\mathcal{S}}| \leqslant \begin{cases} c'''\varepsilon^{-d} & (d-1)p < d, \\ c'''\varepsilon^{-d}\log 1/\varepsilon & (d-1)p = d, \\ c'''\varepsilon^{-(\beta+1)(d-1)/\beta} & (d-1)p > d. \end{cases} \tag{6}$$

3.4. *Proof of Proposition 3.1*

Combining (4) and (6), we have

$$\log N_{[\,]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_p) \leqslant \begin{cases} c\varepsilon^{-d} & (d-1)p < d, \\ c\varepsilon^{-d}(\log 1/\varepsilon)^{1+d/p} & (d-1)p = d, \\ c\varepsilon^{-(d-1)p} & (d-1)p > d, \end{cases}$$

for all $\varepsilon = 2^{-n}$, $n \in \mathbb{N}$. The monotonicity of bracketing numbers implies that Proposition 3.1 holds for all $0 < \varepsilon < 1$.

3.5. *Proof of Corollary 1.3*

Suppose that $[l_i, u_i]$, $i = 1, \ldots, M = N_{[\,]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_p)$ is a collection of ε-brackets for $\mathcal{F}_d$ with respect to $\|\cdot\|_p$. The size of these brackets for $\|\cdot\|_{Q,p}$ is (with λ denoting Lebesgue measure and q the density of Q)

$$\|u_i - l_i\|_{Q,p} = \left\{\int |u_i - l_i|^p q\, d\lambda\right\}^{1/p} \leqslant \|q\|_\infty^{1/p}\varepsilon \equiv \delta,$$

so $\{[l_i, u_i], i = 1, \ldots, M\}$ forms a collection of δ-brackets for $\|\cdot\|_{Q,p}$ and it follows that:

$$\log N_{[\,]}(\varepsilon\|q\|_\infty^{1/p}, \mathcal{F}_d, \|\cdot\|_{Q,p}) \leqslant \log N_{[\,]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_p) \leqslant c_2\varepsilon^{-\alpha}.$$

This yields

$$\log N_{[\,]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_{Q,p}) \leqslant c_2\|q\|_\infty^{\alpha/p}\varepsilon^{-\alpha} = C_2\varepsilon^{-\alpha}$$

with $C_2 \equiv c_2\|q\|_\infty^{\alpha/p}$.

4. Critical case

We believe that the logarithmic factor in Theorem 1.1 is not needed. In this section, we prove that if we only consider the regular entropy, then when $(d, p) \neq (2, 2)$, the logarithmic factor can indeed be removed.

Theorem 4.1. *For* $(d, p) \neq (2, 2)$, *there exist constants* c_1, c_2 *depending only on p and d such that*,

$$c_1\varepsilon^{-\alpha} \leqslant \log N(\varepsilon, \mathcal{F}_d, \|\cdot\|_p) \leqslant c_2\varepsilon^{-\alpha},$$

where $\alpha = \max\{d, (d-1)p\}$.

Proof. In view of Theorem 1.1, it remains to show the upper bound for the case $(d-1)p = d$, $d > 2$. Let

$$T = \{1_A : A = \{(x_1, x_2, \ldots, x_d) : f(x_1, x_2, \ldots, x_d) \leqslant \lambda\}, 0 \leqslant \lambda \leqslant 1, f \in \mathcal{F}_d\}.$$

Then clearly $\mathcal{F}_d$ is the closed convex hull of T, that is $\mathcal{F}_d = \mathrm{conv}(T)$.

For any $1_A \in T$, there exists $f \in \mathcal{F}_d$, and $0 \leqslant \lambda \leqslant 1$ such that

$$A = \{(x_1, \ldots, x_d) : f(x_1, \ldots, x_d) \leqslant \lambda\}.$$

By otherwise changing variable $t_i = 1 - x_i$, we can assume that f is non-decreasing with respect to every variable x_i, $1 \leqslant i \leqslant d$. Define f_A on $[0, 1]^{d-1}$ as follows:

$$f_A(x_1, x_2, \ldots, x_{d-1}) = \begin{cases} \max\{t : (x_1, \ldots, x_{d-1}, t) \in A\} & \text{if } \{t : (x_1, \ldots, x_{d-1}, t) \in A\} \neq \emptyset, \\ 0 & \text{if } \{t : (x_1, \ldots, x_{d-1}, t) \in A\} = \emptyset. \end{cases}$$

It is easy to check that $f_A \in \mathcal{F}_{d-1}$. Furthermore, for all $1_A, 1_B \in T$, $\|1_A - 1_B\|_p = \|f_A - f_B\|_1^{1/p}$. Thus,

$$N_{[\,]}(\varepsilon, T, \|\cdot\|_p) = N_{[\,]}(\varepsilon^p, \mathcal{F}_{d-1}, \|\cdot\|_1).$$

Therefore, by applying Proposition 3.1 for $\mathcal{F}_{d-1}$ with $p = 1$, we have

$$\log N(\varepsilon, T, \|\cdot\|_p) \leqslant \log N_{[\,]}(\varepsilon, T, \|\cdot\|_p) \leqslant c\varepsilon^{-(d-1)p}.$$

Recall a general theorem of [10] (see also [9]) that

$$\log N(\varepsilon, \mathrm{conv}(S)) = O(\varepsilon^{-\sigma})$$

whenever $\log N(\varepsilon, S) = O(\varepsilon^{-\sigma})$ for $\sigma > 2$. Applying these results we obtain

$$\log N(\varepsilon, \mathcal{F}_d, \|\cdot\|_p) = \log N(\varepsilon, \mathrm{conv}(T), \|\cdot\|_p) \leqslant c\varepsilon^{-(d-1)p},$$

for $(d-1)p = d > 2$. $\square$

When $(p, d) = (2, 2)$, we have $(d-1)p = 2$. It was proved in [14] that

$$\log N(\varepsilon, \mathrm{conv}(S)) = O(\varepsilon^{-2}(\log 1/\varepsilon)^2)$$

whenever $\log N(\varepsilon, S) = O(\varepsilon^{-2})$, and there are sets S such that $\log N(\varepsilon, \mathrm{conv}(S)) \geqslant \varepsilon^{-2}(\log 1/\varepsilon)^2$ while $\log N(\varepsilon, S) = O(\varepsilon^{-2})$. Note that the bound $O(\varepsilon^{-2}(\log 1/\varepsilon)^2)$ is exactly the bound we obtained earlier using a direct construction. Thus, in the case $p = d = 2$, using convex hulls does not improve the estimate.

5. Rates of convergence for the MLE of a block decreasing density

Biau and Devroye [1] showed that the minimax rate of convergence for estimating a bounded block decreasing density with L_1 risk is $n^{1/(2+d)}$, and constructed histogram estimators that attain this rate. Here is a more precise description of their result. Let $\mathcal{F}_B$ denote the class of all block decreasing densities on the unit cube $[0, 1]^d$ bounded by B. Define the risk of the estimator $\widehat{f_n}$ when the true density is $f \in \mathcal{B}$ by

$$R(\widehat{f_n}, f) = \mathbb{E}_f \left\{ \int_{\mathbb{R}^d} |\widehat{f_n}(x) - f(x)|\, dx \right\},$$

and the maximum (or "worst case") risk by

$$\mathcal{R}(\widehat{f_n}, \mathcal{F}_B) = \sup_{f \in \mathcal{F}_B} R(\widehat{f_n}, f).$$

The *minimax risk* is $\mathcal{R}_n(\mathcal{F}_B) = \inf_{\widehat{f_n}} \mathcal{R}(\widehat{f_n}, \mathcal{F}_B)$. Biau and Devroye [1] showed that for some constants C_1 and C_2,

$$\mathcal{R}_n(\mathcal{F}_B) \geqslant C_2 \left(\frac{C_1 S^d}{n} \right)^{1/(d+2)},$$

where $S \equiv \log(1 + B)$. The resulting minimax lower bound rate of convergence is $r_n^{mmlb} = n^{1/(2+d)} = n^{\gamma/(2\gamma+1)}$, where $1/\gamma = d$. Biau and Devroye [1] also constructed generalizations of the histogram estimators of Birge [4] which achieve this rate of convergence.

In the case d=1, the MLE of a decreasing density on $[0, M]$ is well-known to be $n^{1/3}$-consistent with respect to Hellinger and L_1 metrics: see Birgé [2,3,5]. In case $d \geqslant 2$, the rate of convergence of the MLE of a block decreasing density with respect to L_1 has been obtained by Polonik [15, Theorem 3.4, p. 69], shows that the "silhouette" converges with respect to L_1 in a general setting of multivariate monotone function estimation, while [16, Theorem 2.3, p. 1862], shows that the silhouette is the MLE for the class of block decreasing densities (which Polonik calls "doubly monotone" densities in the case $d = 2$). By the connection between monotone functions with lower layers and the entropy bounds of [12], it follows that for a bounded block decreasing density f the bracketing entropy hypothesis (3.3) of [15, p. 69], holds with $r = d - 1$, and hence [15, Theorem 3.4] yields (upon noting that Polonik's $\Psi(u) \leqslant u$ as $u \to 0$ for f with support contained in $[0, 1]^d$)

$$\|\widehat{f_n} - f\|_1 = \begin{cases} O_p(n^{-1/4} \log n), & d = 2, \\ O_p(n^{-1/2d}), & d > 2. \end{cases} \tag{7}$$

To relate this result to the Hellinger metric, recall that the L_1 and Hellinger metrics are well-known to satisfy

$$h^2(p, q) \leqslant \|p - q\|_1 \leqslant 2h(p, q);$$

here, the Hellinger distance $h(P, Q)$ is given by $h^2(p, q) = \int [\sqrt{p} - \sqrt{q}]^2\, d\mu$, where μ is any measure dominating both P and Q, p and q are, respectively, the densities of P and Q with respect to μ. Thus (7) combined with the left inequality in the last display only yields $h(\widehat{f_n}, f) = O_p(n^{-1/4d})$ when $d > 2$ and $h(\widehat{f_n}, f) = O_p(n^{-1/8}\sqrt{\log n})$ when $d = 2$, whereas it is known that the MLE converges at the rate $n^{1/3}$ for both the L_1 and Hellinger distance when $d = 1$. Thus, it seems worthwhile to see what the entropy bounds of Section 1 imply concerning the convergence rate of the MLE when $d \geqslant 2$.

It is known from Birgé and Massart [6] (see also [18, pp. 326–327], together with Theorem 3.4.1, p. 322) that MLEs have a rate of convergence of at least $r_n^{\text{mle}} = n^{\gamma/2}$ when the bracketing entropy with respect to the Hellinger metric h of the class of densities $\mathcal{P}$ satisfies

$$\log N_{[\,]}(\varepsilon, \mathcal{P}, h) \leqslant \frac{K}{\varepsilon^{1/\gamma}}, \quad \varepsilon > 0 \tag{8}$$

with $\gamma < \frac{1}{2}$. From the results of [1] it might be guessed that (8) holds for $\mathcal{P} = \mathcal{F}_B$ with $1/\gamma = d$, and this would lead to the rate of convergence $r_n = n^{1/(2d)}$ for the MLE when $d \geqslant 2$. Our theorem 1.1 suggests that the rate of the convergence of the MLE (with respect to Hellinger distance) is still slower than this for $d > 2$, as is shown in the following proposition. We suppose that $X_1, \ldots, X_n$ are i.i.d. $f \in \mathcal{F}_B$.

Proposition 5.1. *Suppose that $\widehat{f}_n$ is the MLE of a block decreasing density f on $[0, 1]^d$. Then*

$$h(\widehat{f}_n, f) = \begin{cases} O_p(n^{-\frac{1}{4(d-1)}}) & \text{if } d \geqslant 3, \\ O_p(n^{-1/4} \log n) & \text{if } d = 2. \end{cases} \tag{9}$$

Remark 5.2. Although we do not yet have lower bounds, the tightness of the entropy bounds in Theorem 1.1 together with the "phase change" in the entropies between $p = 1$ and 2 leads us to conjecture that $n^{1/2d}$ and $n^{1/4(d-1)}$ are the exact rates of convergence of the MLEs of a block-decreasing density in L_1 and Hellinger metrics, respectively, when $d \geqslant 3$.

Proof. We use the results of Birgé and Massart [6] as presented in [18, Section 3.4]. From Theorem 3.4.1, p. 322, with Θ_n taken to be

$$\mathcal{P} = \{p \text{ a block-decreasing density on } [0, 1]^d \text{ bounded by } B\}$$

it follows that we need to establish the inequalities of the first display of p. 323. These follow from Theorem 3.4.4, p. 327, for the Hellinger distance h by choosing $p_n = p_0$ and taking $\mathcal{P}_n = \mathcal{P}$: the resulting bound for $\mathbb{E}_{P_0} \|\mathbb{G}_n\|_{\mathcal{M}_\delta}$ with

$$\mathcal{M}_\delta = \left\{ m_p = \log \frac{p + p_0}{p_0} : \ p \in \mathcal{P} \right\}$$

is of the form

$$\tilde{J}_{[\,]}(\delta, \mathcal{P}, h) \left(1 + \frac{\tilde{J}_{[\,]}(\delta, \mathcal{P}, h)}{\delta^2 \sqrt{n}} \right) \equiv \phi_n(\delta), \tag{10}$$

where

$$\tilde{J}_{[\,]}(\delta, \mathcal{P}, h) = \int_{c\delta^2}^{\delta} \sqrt{1 + \log N_{[\,]}(\varepsilon, \mathcal{P}, h)}\, d\varepsilon$$

in view of the discussion on p. 326 and [6, Theorem 1, p. 118]. Since $\sqrt{p}$ is block-decreasing with bound $\sqrt{B}$ if p is block-decreasing with bound B, it follows that

$$\log N_{[\,]}(\varepsilon, \mathcal{P}, h) = \log N_{[\,]}(\varepsilon, \mathcal{P}^{1/2}, \|\cdot\|_2) = \log N_{[\,]}(\varepsilon/\sqrt{B}, \mathcal{P}^{1/2}/\sqrt{B}, \|\cdot\|_2),$$

where $\|\cdot\|_2$ is the L_2 norm (with respect to Lebesgue measure λ) and where $\mathcal{P}^{1/2}$ is the class of block-decreasing functions with bound $\sqrt{B}$, and hence $\mathcal{P}^{1/2}/\sqrt{B}$ is the class of block-decreasing functions with bound 1. Thus, for $d \geqslant 3$ we calculate, using Theorem 1.1 with $p = 2$,

$$\begin{aligned}
\tilde{J}_{[\,]}(\delta, \mathcal{P}, h) &= \int_{c\delta^2}^{\delta} \sqrt{1 + \log\, N_{[\,]}(\varepsilon, \mathcal{P}, h)}\, d\varepsilon \\
&= \int_{c\delta^2}^{\delta} \sqrt{1 + \log\, N_{[\,]}(\varepsilon/\sqrt{B}, \mathcal{P}^{1/2}/\sqrt{B}, \|\cdot\|_2)}\, d\varepsilon \\
&\leqslant \begin{cases} \int_{c\delta^2}^{\delta} \sqrt{1 + c_2 B^{d-1}} \varepsilon^{-2(d-1)}\, d\varepsilon, & d > 2, \\ \int_{c\delta^2}^{\delta} \sqrt{1 + c_2 B\varepsilon^{-2}(\log\, 1/\varepsilon)^2}\, d\varepsilon, & d = 2, \end{cases} \\
&\lesssim \begin{cases} \delta^{-2(d-2)}, & d > 2, \\ (\log\, 1/\delta)^2, & d = 2, \end{cases}
\end{aligned}$$

where $f(x) \lesssim g(x)$ means $f(x) \leqslant Kg(x)$ for some constant K. Plugging this into (10) yields

$$\phi_n(\delta) = \delta^{-2(d-2)} \left(1 + \frac{\delta^{-2(d-2)}}{\delta^2\sqrt{n}}\right) \quad \text{for } d > 2,$$

$$\phi_n(\delta) = \log(1/\delta)^2 \left(1 + \frac{\log(1/\delta)^2}{\delta^2\sqrt{n}}\right) \quad \text{for } d = 2.$$

It is easily verified that when $d > 2$, $r_n^2 \phi_n(1/r_n) \lesssim \sqrt{n}$ if $r_n = n^{\frac{1}{4(d-1)}}$. When $d = 2$, $r_n^2 \phi_n(1/r_n) \lesssim \sqrt{n}$ if $r_n = n^{\frac{1}{4}} / \log\, n$. Thus, the rate of convergence of the MLE is at least $n^{\frac{1}{4(d-1)}}$ for $d > 2$, and $n^{\frac{1}{4}} / \log\, n$ for $d = 2$. $\square$

Acknowledgment

The authors thank the referee and the editor for several helpful suggestions.

References

[1] G. Biau, L. Devroye, On the risk of estimates for block decreasing densities, J. Multivariate Anal. 86 (2003) 143–165.
[2] L. Birgé, On estimating a density using Hellinger distance and some other strange facts, Probab. Theory Related Fields 71 (1986) 271–291.
[3] L. Birgé, Estimating a density under order restrictions, nonasymptotic minimax risk, Ann. Statist. 15 (1987) 995–1012.
[4] L. Birgé, On the risk of histograms for estimating decreasing densities, Ann. Statist. 15 (1987) 1013–1022.
[5] L. Birgé, The Grenander estimator: a nonasymptotic approach, Ann. Statist. 17 (1989) 1532–1549.
[6] L. Birgé, P. Massart, Rates of convergence for minimum contrast estimators, Probab. Theory Related Fields 97 (1993) 113–150.
[7] L. Birgé, P. Massart, An adaptive compression algorithm in Besov spaces, Constr. Approx. 16 (1) (2000) 1–36.
[8] R. Blei, F. Gao, W. Li, Metric entropy of high dimensional distributions, P. Am. Math. Soc., to appear.
[9] B. Carl, Metric entropy of convex hulls in Hilbert spaces, Bull. London Math. Soc. 29 (1997) 452–458.
[10] B. Carl, I. Kyrezi, A. Pajor, Metric entropy of convex hulls in Banach spaces, J. London Math. Soc. 60 (1999) 871–896.

[11] R.M. Dudley, A course on empirical processes, Lecture Notes in Mathematics, vol. 1097, Springer, Berlin, 1984.
[12] R.M. Dudley, Uniform Central Limit Theorems, Cambridge University Press, Cambridge, England, 1999.
[13] D.E. Edmunds, H. Triebel, Function Spaces, Entropy Numbers and Differential Operators, Cambridge University Press, Cambridge, England, 1996.
[14] F. Gao, Entropy of absolute convex hulls in Hilbert spaces, Bull. London Math. Soc. 36 (2004) 460–468.
[15] W. Polonik, Density estimation under qualitative assumptions in higher dimensions, J. Multivariate Anal. 55 (1995) 61–81.
[16] W. Polonik, The silhouette, concentration functions, and ML-density estimation under order restrictions, Ann. Statist. 26 (1998) 1857–1877.
[17] M. Rudelson, R. Vershynin, Combinatorics of random processes and sections of convex bodies, Annals of Math. 164 (2006) 603–648.
[18] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes, Springer, New York, 1996.
[19] W.P. Ziemer, Weakly Differentiable Functions, Springer, New York, 1989.