

Entropy Estimate for High Dimensional Monotonic Functions

FUCHANG GAO^{*†}
Department of Mathematics
University of Idaho

JON A. WELLNER[‡]
Department of Statistics
University of Washington

December 29, 2005

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.

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

^{*}Corresponding author. Department of Mathematics, P.O. Box 441103, University of Idaho, Moscow, ID 83844-1103. Email: fuchang@uidaho.edu. Phone: 1-208-885-5274. Fax: 1-208-885-5843.

[†]Supported in part by NSF Grant DMS-0405855.

[‡]Supported in part by NSF Grant DMS-0503822.

1 Introduction

Shape constrained functions appear very commonly in nonparametric estimation in statistics via renewal theory and mixing of uniform distributions. A class of multivariate functions of interests in applications is the class of “block-decreasing” densities; see e.g. Polonik [11], [12], and Biau and Devroye [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 nonparametric estimators such as the Maximum Likelihood Estimator (MLE) is determined by 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 entropy $\log N(\varepsilon, \mathcal{F}_d, \|\cdot\|_p)$ and the bracketing entropy $\log N_{[]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_p)$, where $N(\varepsilon, \mathcal{F}_d, \|\cdot\|_p)$ and $N_{[]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_p)$ are defined as follows:

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

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

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

where

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

The new bracketing entropy bounds have implications for the rate of convergence of the Maximum Likelihood Estimator of a “block decreasing” density as will be shown in section 5.

Our main result is the following

Theorem 1.1. For $p \geq 1$, there exist constants c_1 and c_2 depending only on p , such that if $(d-1)p \neq d$, then

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

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

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

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 entropy under the L^p norm, provided $(d, p) \neq (2, 2)$.

It should be pointed out that 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. [13], Theorem 2.75, page 159. So, in some sense, the results in this paper generalize the known results for $d = 1$. It should also be noted that when $d > 1$, \mathcal{F}_d is a much larger class than that of d -dimensional probability distributions. Indeed, Blei, Gao and Li [7] recently proved that under the L^2 norm, the metric entropy of the class \mathcal{D}_d of d -dimensional probability distributions satisfies

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

for $d > 2$, and

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

for $d = 2$.

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 $p \geq 1$, there exists a constant $c_1 > 0$ such that

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

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

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

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

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

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

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

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

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

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

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 $\prod_{i=1}^d (k_i \varepsilon, k_i \varepsilon + \varepsilon)^d$ that satisfies $k_1 + k_2 + \dots + k_d = \varepsilon^{-1}$, $k_1, k_2, \dots, k_d \geq 0$, 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, \dots, q_m , such that $|q_i - q_j| = 1$ on at least $c\varepsilon^{1-d}/2^4$ cubes, $i \neq j$. Thus,

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

This implies that

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

which further implies

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

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 $p \geq 1$, $p \neq d/(d-1)$, there exists a constant $c_2 > 0$ such that

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

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

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

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 $K^{-l} \leq \varepsilon < K^{-l+1}$.

For each $f \in \mathcal{F}_d$, we construct \underline{f} and \bar{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 $\omega(f, I_0) \leq K\varepsilon$. 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 $\omega(f, I_i) \leq K^{i+1}\varepsilon$, 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 $2^{-i}\varepsilon$, $0 \leq i < l$, we define

$$\underline{f} = K^{i+1}\varepsilon \left[\frac{\inf_{x \in I} f(x)}{K^{i+1}\varepsilon} \right], \quad \bar{f} = K^{i+1}\varepsilon \left[\frac{\sup_{x \in I} f(x)}{K^{i+1}\varepsilon} \right].$$

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

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

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

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 \bar{f} , we have $s_i + n_i = 2^d n_{i-1}$. In particular, $s_i \leq 2^d n_{i-1}$.

Now we try to estimate n_{i-1} for $i \geq 1$. 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 $1 \leq j \leq d$, 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

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

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

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

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

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

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

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

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

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

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 $\frac{1}{2K} \leq K^{-p}$. Therefore

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

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

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,

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

for some constants $c, c' > 0$ depending only on p and d , where in the third and fourth inequalities we used the fact $1 \leq K^l \varepsilon < K$ 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

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

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

Summarizing, we obtain that

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

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

We derive the upper bound for $|\bar{\mathcal{S}}|$. 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 $1 \leq i < l$, 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 $0 \leq i < l$, 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 \bar{f} on these r_j cubes are in monotonic order, and are all chosen from $0, K^i \varepsilon, 2K^i \varepsilon, \dots, mK^i \varepsilon$, where $m = \lfloor K^{-i} \varepsilon^{-1} \rfloor$, the number of different ways of assigning values of \bar{f} on these r_j cubes is bounded by

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

Thus, the number of different ways to assign the values of \bar{f} on the s_i selected cubes of side-

length $2^{-i}\varepsilon$ is bounded by

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

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

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

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

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

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

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

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

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

3.4 Proof of Proposition 3.1

Combining (4) and (6), we have

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

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

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,

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

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

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

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

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

$$A = \{(x_1, \dots, x_d) : f(x_1, \dots, x_d) \leq \lambda\}.$$

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

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

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

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

Recall a general theorem of [9] (see also [8]) that

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

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

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

for $(d-1)p = d > 2$. □

When $(p, d) = (2, 2)$, we have $(d-1)p = 2$. It was proved in [10] that

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

whenever $\log N(\varepsilon, S) = O(\varepsilon^{-2})$, and in general, this cannot be improved. Note that this bound is exactly the bound we obtained earlier using a direct construction. Thus, in this case, using convex hulls does not improve the estimate.

5 Rates of convergence for the Maximum Likelihood Estimator 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 \hat{f}_n when the true density is $f \in \mathcal{B}$ by

$$R(\hat{f}_n, f) = E_f \left\{ \int_{\mathbb{R}^d} |\hat{f}_n(x) - f(x)| dx \right\},$$

and the maximum (or “worst case”) risk by

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

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

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

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 $1/\gamma = d$. [1] also constructed generalizations of the histogram estimators of Birgé [4] which achieve this rate of convergence.

The MLE of a decreasing density on $[0, M]$ is well known to be $n^{1/3}$ with respect to Hellinger and L_1 metrics: see Birgé [2], [3], [5]. Although the MLE of a block decreasing density has been initiated by Polonik [12], the rate of convergence of the MLE in this setting with respect to Hellinger or L_1 metrics is apparently unknown for $d \geq 2$. It is known from Birgé and Massart [6] (see also [13], pages 326-327 together with Theorem 3.4.1, page 322) that maximum likelihood estimators have a rate of convergence of at least $r_n^{mle} = n^{\gamma/2}$ when the bracketing entropy with respect to the Hellinger metric h of the class of densities \mathcal{P} satisfies

$$(7) \quad \log N_{[]}(\epsilon, \mathcal{P}, h) \leq \frac{K}{\epsilon^{1/\gamma}}, \quad \epsilon > 0$$

with $\gamma < 1/2$; 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 and p, q are the densities of P, Q with respect to μ . From the results of [1] it might be guessed that (7) 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 $d \geq 2$. 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, \dots, X_n are i.i.d. $f \in \mathcal{F}_B$.

Proposition 5.1. Suppose that \hat{f}_n is the MLE of a block decreasing density f on $[0, 1]^d$. Then if $d \geq 3$

$$(8) \quad n^{\frac{1}{4(d-1)}} h(\hat{f}_n, f) = O_p(1).$$

If $d = 2$, then

$$(9) \quad \frac{n^{1/4}}{\log n} h(\hat{f}_n, f) = O_p(1).$$

Proof. We use the results of Birgé and Massart [6] as presented in section 3.4 of [13]. From Theorem 3.4.1, page 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 page 323. These follow from Theorem 3.4.4, page 327, for the Hellinger distance h by choosing $p_n = p_0$ and taking $\mathcal{P}_n = \mathcal{P}$: the resulting bound for $E_{P_0} \|\mathbb{G}_n\|_{\mathcal{M}_\delta}$ with

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

is of the form

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

where

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

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

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

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 $d \geq 3$ we calculate, using Theorem 1.1 with $p = 2$,

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

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

$$\begin{aligned} \phi_n(\delta) &= \delta^{-2(d-2)} \left(1 + \frac{\delta^{-2(d-2)}}{\delta^2 \sqrt{n}} \right) \quad \text{for } d > 2, \\ \phi_n(\delta) &= (\log(1/\delta))^2 \left(1 + \frac{(\log(1/\delta))^2}{\delta^2 \sqrt{n}} \right) \quad \text{for } d = 2. \end{aligned}$$

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

References

- [1] Biau, G. and Devroye, L. (2003). On the risk of estimates for block decreasing densities. *J. Mult. Anal.* **86**, 143 - 165.
- [2] Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Related Fields* **71**, 271 - 291.
- [3] Birgé, L. (1987a). Estimating a density under order restrictions. nonasymptotic minimax risk. *Ann. Statist.* **15**, 995 - 1012.
- [4] Birgé, L. (1987b). On the risk of histograms for estimating decreasing densities. *Ann. Statist.* **15**, 1013 - 1022.
- [5] Birgé, L. (1989). The Grenander estimator: a nonasymptotic approach. *Ann. Statist.* **17**, 1532-1549.
- [6] Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory and Related Fields* **97**, 113 - 150.
- [7] Blei, R., Gao, F., and Li, W. (2005). Metric entropy of high dimensional distributions and small deviation probability of Brownian sheets. (preprint)
- [8] Carl, B. (1997). Metric entropy of convex hulls in Hilbert spaces. *Bull. London Math. Soc.* **29**, 452 - 458.
- [9] Carl, B., Kyrezi, I., and Pajor, A. (1999). Metric entropy of convex hulls in Banach spaces. *J. London Math. Soc.* **60**, 871 - 896.
- [10] Gao, F. (2004). Entropy of absolute convex hulls in Hilbert spaces. *Bull. London Math. Soc.* **36** (2004), 460 - 468.
- [11] Polonik, W. (1995). Density estimation under qualitative assumptions in higher dimensions. *J. Multivariate Analysis* **55**, 61 - 81.
- [12] Polonik, W. (1998). The silhouette, concentration functions, and ML-density estimation under order restrictions. *Ann. Statist.* **26**, 1857 - 1877.
- [13] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.