Covering numbers, Vapnik-Červonenkis classes and bounds for the star-discrepancy

Aicke Hinrichs*

December 2, 2003

Abstract

We show new lower bounds for the star-discrepancy and its inverse for subsets of the unit cube. They are polynomial in the quotient d/n of the number n of sample points and the dimension d. They provide the best known lower bounds for n not too large compared with d.

1 Introduction

The star-discrepancy of a set is one of the main tools to estimate the worst case error of multivariate integration for certain classes of functions. There are rather accurate bounds on the best possible star-discrepancy of an n-point set in the d-dimensional unit cube $[0,1]^d$ for fixed dimension d and, compared with d, very large n (typically exponential in d), see e.g the monographs of H. Niederreiter [Nie92] and M. Drmota and R. F. Tichy [DT97]. For some applications the dimension d may be so large that it becomes impossible to use enough points such that these bounds give reasonable error estimates. The question how the discrepancy depends on the dimension d then turns out to be a critical issue.

S. Heinrich, E. Novak, G. W. Wasilkowski and H. Woźniakowski recently studied this question in [HNWW01]. They prove an upper bound on the star discrepancy that shows that it depends only polynomially on d/n. They also show a lower bound which is rather far from the upper bound as it depends exponentially on d/n. It is the purpose of this paper to show that the lower bound can be improved to a polynomial behavior in d/n.

To state problems and results precisely let us introduce the necessary notation. Cardinality of a finite set A and Lebesgue measure of a measurable subset B of $\mathbb{R}^d$ are denoted by $|A|$ and $|B|$, respectively. Let T be a finite subset of the d-dimensional unit cube $I^d = [0,1]^d$ with $|T| = n$. Given $x \in I^d$, we consider the box $C_x = \{y \in I^d : 0 \le y_i < x_i \text{ for } i = 1, \dots, d\}$. The discrepancy of T in the box C_x is given by

$$D(T,x) = |C_x| - n^{-1}|T \cap C_x|.$$

*Research of the author is supported by the DFG Emmy-Noether grant Hi 584/2-3.

The star-discrepancy of T is then defined as

$$D_\infty^*(T) = \sup_{x \in I^d} |D(T,x)|.$$

Let

$$D_\infty^*(n,d) = \inf\{D_\infty^*(T) \,:\, T \subset I^d, |T| = n\}$$

be the minimal star-discrepancy of an n-point subset in dimension d. Our results are sometimes better expressed via the inverse function

$$N_\infty^*(d,\varepsilon) = \min\{n \,:\, D_\infty^*(n,d) \le \varepsilon\} = \min\{|T| \,:\, T \subset I^d, D_\infty^*(T) \le \varepsilon\}$$

for $\varepsilon \in (0,1)$.

The main result in [HNWW01] states that there exist constants $c_1, c_2, \varepsilon_0 > 0$ such that for all $0 < \varepsilon \le \varepsilon_0$

$$c_1 d \log(1/\varepsilon) \le N_\infty^*(d,\varepsilon) \le c_2 d/\varepsilon^2, \tag{1}$$

or, in terms of the discrepancy, there exist constants $c_1, c_2, \varepsilon_0 > 0$ such that for all d, n

$$\min(\varepsilon_0, e^{-c_1 n/d}) \le D_\infty^*(n,d) \le c_2\sqrt{d/n}. \tag{2}$$

In [Hei03], the problem was raised to narrow the considerable gap between lower and upper bounds in (1) in terms of the dependence on ε. We prove here

Theorem 1. *There exist constants $c, \varepsilon_0 > 0$ such that*

$$N_\infty^*(d,\varepsilon) \ge cd/\varepsilon \quad \text{for } 0 < \varepsilon < \varepsilon_0 \tag{3}$$

and

$$D_\infty^*(n,d) \ge \min(\varepsilon_0, cd/n). \tag{4}$$

In fact, our proof shows this lower bound for the discrepancy defined with arbitrary coefficients instead of the uniform weights $1/n$. Let $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and $T = \{t_1, \ldots, t_n\} \subset I^d$. Then we consider the discrepancy function with weight a given by

$$D(T,a,x) = |C_x| - \sum_{i=1}^n a_i \mathbf{1}_{C_x}(t_i),$$

where $\mathbf{1}_C$ denotes the indicator function of the set C. Obviously, $D(T,x) = D(T,a,x)$ for $a = (1/n, \ldots, 1/n)$. Furthermore, we define as in the uniformly weighted case

$$D_\infty^*(T,a) = \sup_{x \in I^d} |D(T,a,x)|.$$

Finally, let

$$D_\infty(n,d) = \inf\{D_\infty(T,a) \,:\, T \subset I^d, |T| = n, a \in \mathbb{R}^n\}$$

be the minimal weighted star-discrepancy of an n-point subset in dimension d and

$$N_\infty(d,\varepsilon) = \min\{n \,:\, D_\infty(n,d) \le \varepsilon\}$$

its inverse function. Obviously, $D_\infty(n,d) \le D_\infty^*(n,d)$ and $N_\infty(n,d) \le N_\infty^*(n,d)$. Then Theorem 1 is an immediate consequence of

Theorem 2. *There exist constants $c, \varepsilon_0 > 0$ such that*

$$N_\infty(d, \varepsilon) \geq cd/\varepsilon \quad \textit{for } 0 < \varepsilon < \varepsilon_0 \tag{5}$$

and

$$D_\infty(n, d) \geq \min(\varepsilon_0, cd/n). \tag{6}$$

The proof of the upper bound in (1) and (2) in [HNWW01] is based on the fact that the class of boxes $(C_x)_{x \in I^d}$ is a Vapnik-Červonenkis class of dimension d. A feature of our proof of the lower bound is that it also uses essentially the VC-property of this class. Let us recall the necessary notions and results.

Let $(X, \mathbb{P})$ be a probability space with probability measure $\mathbb{P}$. A countable family $\mathcal{C}$ of measurable subsets of X is called a Vapnik-Červonenkis class (for short VC-class) if there exists a nonnegative integer v such that

$$|\{A \cap C \,:\, C \in \mathcal{C}\}| < 2^{v+1}$$

for any subset $A \subset X$ with $|A| = v + 1$. The smallest such v is called VC-dimension of $\mathcal{C}$. A basic inequality for any VC-class of dimension v, independently due to N. Sauer [Sau72], S. Shelah [She72], V. N. Vapnik and A. Ya. Červonenkis [VC71], is the estimate

$$|\{A \cap C \,:\, C \in \mathcal{C}\}| \leq \sum_{i=0}^{v} \binom{|A|}{i} \tag{7}$$

for any finite set $A \subset X$.

The discrepancy of an n-element set $T = \{t_1, \ldots, t_n\} \subset X$ with respect to $C \in \mathcal{C}$ and weight $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ is then given as

$$D(T, a, C) = \mathbb{P}(C) - \sum_{i=1}^{n} a_i \mathbf{1}_C(t_i).$$

The discrepancy of T with respect to $\mathcal{C}$ and a is

$$D_\infty^{\mathcal{C}}(T, a) = \sup_{C \in \mathcal{C}} |D(T, a, C)|.$$

Again, we abbreviate $D_\infty^{\mathcal{C}}(T) = D_\infty^{\mathcal{C}}(T, a)$ for the uniform weight $a = (1/n, \ldots, 1/n)$. Observe that the such defined discrepancy with respect to the class of boxes (C_x) in I^d (with x restricted to $I^d \cap \mathbb{Q}^d$) is exactly the star discrepancy defined earlier. In [HNWW01], the upper bounds in (1) and (2) are proved in this more general setting:

Theorem 3. *There is a constant $c > 0$ such that for any VC-class $\mathcal{C}$ of dimension v, any probability $\mathbb{P}$ as above and any $n \in \mathbb{N}$ there exists $T \subset X$ with $|T| = n$ and $D_\infty^{\mathcal{C}}(T) \leq c\sqrt{v/n}$.*

Our lower bound from Theorem 2 also allows such a generalization taking into account the covering numbers of the class $\mathcal{C}$ with respect to the pseudometric $d_{\mathbb{P}}(C_1, C_2) = \mathbb{P}(C_1 \Delta C_2)$. For $\varepsilon > 0$, the covering number $N(M, d, \varepsilon)$ of a pseudometric space (M, d) is the smallest number of (closed) ε-balls $B(x, \varepsilon) = \{y \in M : d(x, y) \leq \varepsilon\}$ that cover M. If there is no finite covering with ε-balls we set $N(M, d, \varepsilon) = \infty$. We will need the obvious fact that if M is equipped with two metrics d_1 and d_2 such that $d_1 \geq d_2$ then $N(M, d_1, \varepsilon) \geq N(M, d_2, \varepsilon)$.

Theorem 4. *Let $\mathcal{C}$ be a VC-class of dimension v which is closed under intersections and let $\mathbb{P}$ be a probability as above. Assume that there exists a constant $\kappa > 0$ such that $N(\mathcal{C}, d_{\mathbb{P}}, \varepsilon) \geq (\kappa\varepsilon)^{-v}$ for all $\varepsilon > 0$. Then there exist constants $c, \varepsilon_0 > 0$ such that for all n, all $T \subset X$ with $|T| = n$ and all $a \in \mathbb{R}^n$*

$$D_\infty^{\mathcal{C}}(T, a) \geq \min(\varepsilon_0, cv/n).$$

Observe that by a result of D. Haussler [Hau95], the required entropy behavior in the assumption of the theorem is essentially the worst possible for any VC-class of dimension v.

2 Proofs

The attentive reader familiar with the lower bound proof of (1) in [HNWW01] will observe that our approach was inspired by that one. We first prove Theorem 4 and then deduce Theorem 2 from it by estimating the relevant covering numbers. We start with two lemmas. We always deal with a v-dimensional VC-class $\mathcal{C}$ of measurable subsets of a ground set X equipped with a probability $\mathbb{P}$.

Lemma 5. *Assume that $T \subset X$ with $|T| = n$ and $\varepsilon > 0$ satisfy*

$$\sum_{i=0}^{v} \binom{n}{i} < N(\mathcal{C}, d_{\mathbb{P}}, \varepsilon).$$

Then there exist $C_1, C_2 \in \mathcal{C}$ such that $\mathbb{P}(C_1 \Delta C_2) = d_{\mathbb{P}}(C_1, C_2) > \varepsilon$ and $T \cap C_1 = T \cap C_2$.

Proof. Let $\mathcal{N}$ be a maximal subset of $\mathcal{C}$ such that $d_{\mathbb{P}}(C_1, C_2) > \varepsilon$ for distinct $C_1, C_2 \in \mathcal{N}$. The maximality of $\mathcal{N}$ implies that the closed balls with centers in $\mathcal{N}$ and radius ε cover $\mathcal{C}$. Hence $N(\mathcal{C}, d_{\mathbb{P}}, \varepsilon) \leq |\mathcal{N}|$. Now it follows from (7) and the assumption that

$$|\{T \cap C : C \in \mathcal{C}\}| \leq \sum_{i=0}^{v} \binom{n}{i} < |\mathcal{N}|.$$

Hence there must exist distinct $C_1, C_2 \in \mathcal{N}$ such that $T \cap C_1 = T \cap C_2$. $\square$

Lemma 6. *Let $T \subset X$, $\varepsilon > 0$, and $C_1, C_2 \in \mathcal{C}$ be such that $\mathbb{P}(C_1 \Delta C_2) = d_{\mathbb{P}}(C_1, C_2) > \varepsilon$ and $T \cap C_1 = T \cap C_2$. If $\mathcal{C}$ is closed under intersections then $D_\infty^{\mathcal{C}}(T, a) \geq \varepsilon/4$ for all $a \in \mathbb{R}^n$.*

Proof. Let $C = C_1 \cap C_2 \in \mathcal{C}$, $n = |T|$, and $S = T \cap C_1 = T \cap C_2 = T \cap C$. Let $\alpha = \sum_{i=1}^n a_i \mathbf{1}_S(t_i)$. Then

$$\begin{aligned} 4D_\infty^{\mathcal{C}}(T, a) &\geq D(T, a, C_1) + D(T, a, C_2) - 2D(T, a, C) \\ &= (\mathbb{P}(C_1) - \alpha) + (\mathbb{P}(C_2) - \alpha) - 2(\mathbb{P}(C) - \alpha) \\ &= \mathbb{P}(C_1) + \mathbb{P}(C_2) - 2\mathbb{P}(C) = \mathbb{P}(C_1 \Delta C_2) > \varepsilon. \end{aligned}$$

□

Proof of Theorem 4. We may assume that $\kappa \geq 1$. Choose $c = 1/4\kappa e$ and assume that $n \geq v$ and $0 < \varepsilon \leq cv/n$. Then

$$\sum_{i=0}^{v} \binom{n}{i} < \left(\frac{en}{v}\right)^v \leq \left(\frac{1}{4\kappa\varepsilon}\right)^v \leq N(\mathcal{C}, d_{\mathbb{P}}, 4\varepsilon).$$

Then Lemma 5 implies that there exist $C_1, C_2 \in \mathcal{C}$ such that $\mathbb{P}(C_1 \Delta C_2) = d_{\mathbb{P}}(C_1, C_2) > 4\varepsilon$ and $T \cap C_1 = T \cap C_2$. Now Lemma 6 shows that $D_\infty^{\mathcal{C}}(T, a) \geq \varepsilon$. Hence

$$D_\infty^{\mathcal{C}}(T, a) \geq cv/n \quad \text{whenever } n = |T| \geq v.$$

If on the other hand $n = |T| < v$, we may choose an extension $S \supset T$ with $|S| = v$. Finally, also extending the weight $a \in \mathbb{R}^n$ to a weight $b \in \mathbb{R}^v$ by letting $b_i = a_i$ for $i = 1, \ldots, v$ and $b_i = 0$ for $i = v+1, \ldots, n$, it follows from what was already shown that

$$D_\infty^{\mathcal{C}}(T, a) = D_\infty^{\mathcal{C}}(S, b) \geq c =: \varepsilon_0.$$

□

Remark. To not clutter up our argument, we did not try to optimize the constants in our results. It is obvious, that better estimates on $\sum_{i=0}^{v} \binom{n}{i}$ for particular choices of n, v lead to better constants.

Proof of Theorem 2. We first derive (6) from Theorem 4. To that end, let $\mathcal{C}$ be the system of all boxes C_x with $x \in I^d \cap \mathbb{Q}^d$ and let $\mathbb{P}$ be Lebesgue measure on I^d. Then $D_\infty^*(T, a) = D_\infty^{\mathcal{C}}(T, a)$ for any finite $T \in I^d$ and $a \in \mathbb{R}^{|T|}$. It is well-known that the class $\mathcal{C}$ has VC-dimension $v = d$, see e.g. [Dud84, Corollary 9.2.15]. So Theorem 4 indeed implies (6) once we show that there is some $\kappa > 0$ such that

$$N(\mathcal{C}, d_{\mathbb{P}}, \varepsilon) \geq (\kappa\varepsilon)^{-d} \quad \text{for all } \varepsilon > 0. \tag{8}$$

This was basically already done in [HNWW01]. Nevertheless we include a short argument here for the convenience of the reader.

Let us transfer the distance $d_{\mathbb{P}}$ to I^d by $d(x,y) := d_{\mathbb{P}}(C_x, C_y)$. We first show that

$$d(x,y) \geq |C_{x\wedge y}|\, \|x-y\|_1 \quad \text{for } x,y \in I^d \tag{9}$$

where $x \wedge y$ is the coordinatewise minimum of x and y and $\|u\|_1 = \sum_{i=1}^d |u_i|$ is the l_1-norm of $u \in \mathbb{R}^d$. Indeed, let $z = x \wedge y$ and estimate

$$\begin{aligned} |C_x \setminus C_y| &= |C_x \setminus C_z| = \prod_{i=1}^d x_i - \prod_{i=1}^d z_i = \sum_{i=1}^d x_1 \dots x_{i-1}(x_i - z_i) z_{i+1} \dots z_d \\ &\geq \prod_{i=1}^d z_i \sum_{i=1}^d (x_i - z_i) = |C_z| \sum_{i=1}^d (x_i - z_i). \end{aligned}$$

Analogously, we obtain

$$|C_y \setminus C_x| \geq |C_z| \sum_{i=1}^d (y_i - z_i).$$

Both inequalities together yield as claimed

$$d(x,y) = |C_x \Delta C_y| \geq |C_z| \sum_{i=1}^d (x_i + y_i - 2z_i) = |C_z| \sum_{i=1}^d |x_i - y_i|.$$

If $x, y \in M := [1 - 1/2d, 1]^d$, it follows from $|C_{x\wedge y}| \geq (1-1/2d)^d \geq 1/2$ that $d(x,y) \geq \|x-y\|_1/2$. Hence a simple volume comparison implies that

$$N(\mathcal{C}, d_{\mathbb{P}}, \varepsilon) = N(I^d, d, \varepsilon) \geq N(M, d, \varepsilon) \geq N(M, \|\,.\,\|_1, 2\varepsilon) \geq \frac{|M|}{(2\varepsilon)^d |B|},$$

where $|B| = 2^d/d!$ is the volume of the l_1^d-unit ball $B = \{x \in \mathbb{R}^d : \|x\|_1 \leq 1\}$. So we arrive at

$$N(\mathcal{C}, d_{\mathbb{P}}, \varepsilon) \geq \frac{d!}{(8\varepsilon d)^d} \geq (8e\varepsilon)^{-d}$$

which is (8) with $\kappa = 8e$.

Finally, inequality (5) is immediate from (6). □

Acknowledgement. The author thanks Erich Novak for introducing him to this problem and for some useful discussions about the contents of this note.

References

[DT97] M. Drmota, R. F. Tichy, *Sequences, discrepancies and applications.* Springer Lecture Notes in Math. **1651**, 1997.

[Dud84] R. M. Dudley, *A course on empirical processes.* In: École d'été de probabilités de Saint-Flour, XII—1982, Springer Lecture Notes in Math. **1097** (1984), 1–142.

[Hau95] D. Haussler, *Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Červonenkis dimension.* J. Combin. Theory Ser. A **69** (1995), 217–232.

[HNWW01] S. Heinrich, E. Novak, G. W. Wasilkowski, H. Woźniakowski, *The inverse of the star-discrepancy depends linearly on the dimension.* Acta Arith. **96** (2001), 279–302.

[Nie92] H. Niederreiter, *Random number generation and quasi-Monte Carlo methods.* CBMS-NSF Regional Conference Series in Applied Mathematics **63**, SIAM, Philadelphia, 1992.

[Hei03] S. Heinrich, *Some open problems concerning the star-discrepancy.* J. Complexity **19** (2003), 416–419.

[Sau72] N. Sauer, *On the density of families of sets.* J. Combinatorial Theory Ser. A **13** (1972), 145–147.

[She72] S. Shelah, *A combinatorial problem; stability and order for models and theories in infinitary languages.* Pacific J. Math. **41** (1972), 247–261.

[VC71] V. N. Vapnik, A. Ya. Červonenkis, *On the uniform convergence of relative frequencies events to their probabilities.* Theor. Probability Appl. **16** (1971), 264–280.

Department of Mathematics, Friedrich-Schiller-University Jena,
Ernst-Abbe-Platz 2, D-07743 Jena, Germany
E-mail: nah@rz.uni-jena.de