Combinatorics, Probability and Computing (2004) 13, 353–360. © 2004 Cambridge University Press
DOI: 10.1017/S0963548303005911 Printed in the United Kingdom

Fat-Shattering of Affine Functions

DON HUSH and CLINT SCOVEL

Modeling, Algorithms and Informatics Group, CCS-3, Mail Stop B265,
Los Alamos National Laboratory, Los Alamos, NM 87545, USA
(e-mail: {dhush,jcs}@lanl.gov)

Received 5 February 2003; revised 17 June 2003

We compute the fat-shattering function and the level fat-shattering function for important classes of affine functions. We observe that the level fat-shattering function and the fat-shattering function are identical for these classes. In addition we observe that the notion that adding the constant term to linear functions increases the dimension by at most 1 is incorrect for fat-shattering and level fat-shattering.

Fat-shattering was introduced by Kearns and Schapire [6] to provide lower bounds on sample complexity for a learning problem. However it appears that fat-shattering is useful more generally. In particular, Alon, Ben-David, Cesa-Bianchi and Haussler [1] have proved a generalization of Sauer's lemma, bounding covering numbers of a function class in terms of its fat-shattering dimension, leading to a new characterization of Glivenko–Cantelli classes. Shawe-Taylor, Bartlett, Williamson and Anthony [7] use the lemma of [1] to provide the first justification for the performance of margin-based classifiers such as Vapnik's support vector machines [8]. Consequently it appears useful to understand the fat-shattering dimension of function classes.

Gurvits [4] investigates linear functions on Banach spaces and provides bounds on the fat-shattering function in terms of Banach space geometry. He demonstrates in general an increase in dimension when going from level fat-shattering to fat-shattering. However, the situation for Hilbert spaces appears to be quite different. Shawe-Taylor, Bartlett, Williamson and Anthony [7] show that when B_R is the ball of radius R in an inner-product space H of dimension n and $\mathscr{F}$ is the class of affine functions $f(x) = \phi \cdot x + c$ with $|\phi|_H \leqslant 1$ and $|c| \leqslant R$ on the domain B_R, then

$$\mathrm{fat}_{\mathscr{F}}(\gamma) \leqslant \min\left(\frac{9R^2}{\gamma^2}, n+1\right) + 1.$$

(We remark that we allow $n = \infty$; in this case we take $n+1 = \infty$.) In this paper we

https://doi.org/10.1017/S0963548303005911 Published online by Cambridge University Press

eliminate the constraint on c and compute the shattering function *exactly*. As a corollary we obtain a simple approximate form for the shattering function which sharpens this bound essentially to an equality

$$\max\left(\frac{R^2}{\gamma^2}, 1\right) \leqslant \mathrm{fat}_{\mathscr{F}}(\gamma) \leqslant \min\left(\frac{R^2}{\gamma^2} + \frac{5}{4}, n+1\right).$$

The proof relies on that used in Hush and Scovel [5] to prove a theorem of Vapnik on the *level* fat-shattering dimension of the affine functions. It is modelled on that found in Bartlett and Shawe-Taylor [2], which they applied to the linear case. Their technique seems to be based on Gurvits [4]. In the process of proof we obtain that the fat-shattering function and the *level* fat-shattering function are identical for the affine functions. Comparing with bounds of Bartlett and Shawe-Taylor [2] for linear functions, we observe that the notion that going from linear to affine functions should add at most 1 to the fat-shattering or the level fat-shattering dimension is incorrect.

Definition. Let $\mathscr{F}$ denote a set of real-valued functions on a set X and consider $\gamma > 0$. We say that a subset $A \subset X$ is γ-shattered by $\mathscr{F}$ if there is a real vector parametrized by A, $\{r_a \in \mathbb{R} : a \in A\}$, such that for all binary vectors b parametrized by A, there is a function f_b such that $f_b(a) \geqslant r_a + \gamma$ when $b_a = 1$ and $f_b(a) \leqslant r_a - \gamma$ otherwise. The fat-shattering dimension of $\mathscr{F}$ at scale γ, $\mathrm{fat}_{\mathscr{F}}(\gamma)$, is the size of the largest subset $A \subset X$ which is γ-shattered by $\mathscr{F}$.

The definition of γ *level* shattering is the same as above but where $r_a = r$ is constant in a.

We now state and prove our main result.

Theorem 1. *Let H denote an inner-product space of dimension n and consider its closed ball $X = B_{R_1} \subset H$ of radius R_1. Let $\mathscr{F}$ denote the class of functions on X defined by $f(x) = \phi \cdot x + c$ with $|\phi|_H \leqslant R_2$ and $c \in \mathbb{R}$. Let*

$$\gamma_k = \frac{1}{\sqrt{k-1}} \quad k \text{ even},$$
$$\gamma_k = \frac{k}{k-1}\frac{1}{\sqrt{k+1}} \quad k \text{ odd}.$$

Then

$$\mathrm{fat}_{\mathscr{F}}(\gamma) = n+1, \quad \frac{\gamma}{R_1 R_2} \leqslant \gamma_{n+1}, \tag{1}$$

$$\mathrm{fat}_{\mathscr{F}}(\gamma) = k, \quad \gamma_{k+1} < \frac{\gamma}{R_1 R_2} \leqslant \gamma_k, \quad 1 \leqslant k \leqslant n+1. \tag{2}$$

When $n = \infty$ the first line (1) never occurs, because $\gamma > 0$.

Proof. We prove the theorem for $R_1 = R_2 = 1$. The general theorem then follows from simple scaling arguments. We first prove that

$$1 \leqslant \mathrm{fat}_{\mathscr{F}}(\gamma) \leqslant n+1. \tag{3}$$

https://doi.org/10.1017/S0963548303005911 Published online by Cambridge University Press

The inequality $1 \leqslant \mathrm{fat}_{\mathscr{F}}(\gamma)$ follows from the fact that c can be any real number. We now assume n is finite, for otherwise there is nothing to prove. Let $A = \{x_1, \ldots, x_k\}$, where $x_i \in X, i = 1, \ldots, k$. From the definition, if A is γ-shattered by $\mathscr{F}$ then there exist k real values $r_i, i = 1, \ldots, k$ such that, for every binary vector $b = (b_1, \ldots, b_k) \in \{-1, 1\}^k$, there is a (ϕ_b, c_b) with $|\phi_b| \leqslant 1$ such that

$$\phi_b \cdot x_i + c_b \geqslant r_i + \gamma, \quad b_i = 1, \tag{4}$$

$$\phi_b \cdot x_i + c_b \leqslant r_i - \gamma, \quad b_i = -1. \tag{5}$$

If we define $\tilde{x} = (x, r)$, $\tilde{x}_i = (x_i, r_i)$, $\tilde{\phi} = (\phi, -1)$ and $\tilde{\phi}_b = (\phi_b, -1)$, then we obtain

$$\tilde{\phi}_b \cdot \tilde{x}_i + c_b \geqslant \gamma, \quad b_i = 1,$$
$$\tilde{\phi}_b \cdot \tilde{x}_i + c_b \leqslant -\gamma, \quad b_i = -1.$$

Since $\gamma > 0$ this implies that

$$\tilde{\phi}_b \cdot \tilde{x}_i + c_b \geqslant 0, \quad b_i = 1, \tag{6}$$

$$\tilde{\phi}_b \cdot \tilde{x}_i + c_b < 0, \quad b_i = -1. \tag{7}$$

The set of functions $\tilde{x} \mapsto \tilde{\phi} \cdot \tilde{x} + c = \phi \cdot x - r + c$, as (ϕ, c) vary, is a subset of an affine space of dimension $\leqslant n + 1$. Applying the following generalization of a theorem of Steele and Dudley (see [3, Theorem 13.9]) to this affine space of functions satisfying (6) and (7) finishes the proof of (3). We note that this proof and therefore the corresponding bound also applies for level fat-shattering.

Lemma 1. *Let $\mathscr{G}$ be an affine space of real functions on a set X with $\dim(\mathscr{G}) = r$. The class of sets*

$$\mathscr{A} = \{\{x : g(x) \geqslant 0\} : g \in \mathscr{G}\}$$

has VC dimension $V_{\mathscr{A}} \leqslant r$.

Proof. The proof follows that of [3, Theorem 13.9]. It suffices to show that no set of size $k > r$ points can be shattered by $\mathscr{A}$. Fix arbitrary points $x_1, \ldots, x_k$ in X and consider the linear mapping L defined by

$$g \mapsto (g(x_1), \ldots, g(x_k)) \in \mathbb{R}^k.$$

Since $\mathscr{G}$ is affine and L is linear, $L\mathscr{G}$ is an affine subspace of $\mathbb{R}^k$ with $\dim(L\mathscr{G}) \leqslant r < k$. If this affine subspace contains the origin, then there is a nonzero vector α orthogonal to $L\mathscr{G}$ with at least one positive value. If we define $S = \{i : \alpha_i \leqslant 0\}$ then the set of points $\{x_i : \alpha_i \leqslant 0\}$ can never equal a set $\{x_i : g(x_i) \geqslant 0\}$ for any g, because the right-hand side of

$$0 = \langle \alpha, Lg \rangle = \sum_i \alpha_i g(x_i) = \sum_S \alpha_i g(x_i) + \sum_{S^c} \alpha_i g(x_i)$$

would be strictly negative, giving a contradiction. Consequently $\mathscr{A}$ cannot shatter $x_1, \ldots, x_k$. Similarly, if the affine subspace $L\mathscr{G}$ does not contain the origin, a vector α in

https://doi.org/10.1017/S0963548303005911 Published online by Cambridge University Press

$L\mathscr{G}$ closest to the origin satisfies

$$\langle \alpha, (Lg - \alpha) \rangle = 0, \quad \forall g \in \mathscr{G},$$

which means that

$$\langle \alpha, Lg \rangle = \sum_i \alpha_i g(x_i) = |\alpha|^2 > 0.$$

If we define $S = \{i : \alpha_i \leqslant 0\}$, then the set of points $\{x_i : \alpha_i \leqslant 0\}$ can never equal a set $\{x_i : g(x_i) \geqslant 0\}$ for any g, because the left-hand side of

$$\sum_i \alpha_i g(x_i) = \sum_S \alpha_i g(x_i) + \sum_{S^c} \alpha_i g(x_i) > 0$$

would be strictly negative, giving a contradiction. Consequently $\mathscr{A}$ cannot shatter $x_1, \ldots, x_k$. □

We now proceed to obtain stronger inequalities.

Lemma 2. *Suppose that $k > 1$ points are contained in the ball of radius 1 in H and are γ fat-shattered by the affine linear functions $\phi \cdot x + c$ where $|\phi| \leqslant 1$ and $c \in \mathbb{R}$. Then*

$$\gamma \leqslant \gamma_k.$$

We note that Hush and Scovel [5] prove the same inequalities of Lemma 2 for level fat-shattering.

Proof. Let the binary vector b determine a nontrivial partition of the index set. We define the weight

$$L(b)_i = \frac{1}{|\{j : b_j = b_i\}|}$$

so that

$$\sum_{i=1}^{k} L(b)_i = 2$$

and

$$\sum_{i=1}^{k} L(b)_i b_i = 0.$$

If the k points x_i are γ shattered, then there is a k-vector r such that, for each binary vector b, there is a (ϕ_b, c_b) that satisfies (4) and (5). These two equations can be written

$$b_i \phi_b \cdot x_i + b_i c_b \geqslant b_i r_i + \gamma \quad i = 1, \ldots, k. \tag{8}$$

Since $L(-b) = L(b)$, the sum $\sum L(-b)_i (-b)_i r_i = -\sum L(b)_i b_i r_i$ is odd with respect to reflection and, among b and $-b$, at least one satisfies

$$\sum L(b)_i b_i r_i \geqslant 0.$$

https://doi.org/10.1017/S0963548303005911 Published online by Cambridge University Press

Consider this choice of b. Since the weights $L(b)_i$ are positive, we multiply equation (8) by $L(b)_i$ and sum to obtain

$$\phi_b \cdot \sum L(b)_i b_i x_i \geqslant 2\gamma,$$

where we have used the fact that $\sum_{i=1}^{k} L(b)_i = 2$ and $\sum_{i=1}^{k} L(b)_i b_i = 0$. Since $|\phi_b| \leqslant 1$, the Cauchy–Schwartz inequality implies that

$$\left| \sum L(b)_i b_i x_i \right| \geqslant 2\gamma.$$

Since $\sum L(-b)_i(-b_i)x_i = -\sum L(b)_i b_i x_i$, this inequality is also true for $-b$, and consequently for all b representing nontrivial partitions.

Hush and Scovel [5] show that, when k is even,

$$E\left(\left| \sum L(b)_i b_i x_i \right|^2 \right) \leqslant \frac{4}{k-1},$$

where E denotes expectation with respect to the uniform distribution over all partitions of the k indices into two $\frac{k}{2}$-sized subsets. They also showed that, when k is odd,

$$E\left(\left| \sum L(b)_i b_i x_i \right|^2 \right) \leqslant \frac{4k^2}{(k-1)^2(k+1)},$$

where E denotes expectation with respect to the uniform distribution over all partitions into a size r subset and a size $r+1$ subset where $r = \frac{k+1}{2}$.

When k is even there must exist a partition b such that

$$\left| \sum L(b)_i b_i x_i \right|^2 \leqslant E\left(\left| \sum L(b)_i b_i x_i \right|^2 \right) \leqslant \frac{4}{k-1},$$

but since we know that

$$\left| \sum L(b)_i b_i x_i \right| \geqslant 2\gamma,$$

we obtain that

$$\gamma \leqslant \frac{1}{\sqrt{k-1}} \quad k \text{ even}.$$

Likewise, when k is odd, there must exist a partition such that

$$\left| \sum L(b)_i b_i x_i \right|^2 \leqslant E\left(\left| \sum L(b)_i b_i x_i \right|^2 \right) \leqslant \frac{4k^2}{(k-1)^2(k+1)},$$

giving the inequality

$$\gamma \leqslant \frac{k}{k-1}\frac{1}{\sqrt{k+1}} \quad k \text{ odd}.$$

The proof of Lemma 2 is finished. □

We now complete the proof of Theorem 1. Since the inequalities of Lemma 2 and the bound (3) apply for level fat-shattering, it is not hard to see that the following argument constructs both the fat-shattering function and the level fat-shattering function

https://doi.org/10.1017/S0963548303005911 Published online by Cambridge University Press

simultaneously, and that they are identical. Vapnik [8] observed that when $k \leqslant n+1$ is finite the regular unit k-simplex achieves

$$\gamma = \gamma_k,$$

so that $\text{fat}_{\mathscr{F}}(\gamma_k) \geqslant k$. Since $\text{fat}_{\mathscr{F}}$ is non-increasing in γ, we obtain that

$$\gamma \leqslant \gamma_k \Rightarrow \text{fat}_{\mathscr{F}}(\gamma) \geqslant k, \quad k \leqslant n+1. \tag{9}$$

The inequalities of Lemma 2 imply that

$$\gamma > \gamma_k \Rightarrow \text{fat}_{\mathscr{F}}(\gamma) < k, \quad 1 < k. \tag{10}$$

Combining (9) and (10) we obtain

$$\gamma_{k+1} < \gamma \leqslant \gamma_k, \quad 1 < k \leqslant n+1 \Rightarrow \text{fat}_{\mathscr{F}}(\gamma) = k \tag{11}$$

and from (9), the monotonicity of $\text{fat}_{\mathscr{F}}$, and the bound (3), we obtain

$$\gamma \leqslant \gamma_{n+1} \Rightarrow \text{fat}_{\mathscr{F}}(\gamma) = n+1,$$

and so obtain equations (1) and (2) restricted to $k > 1$. It follows from (11), the monotonicity of $\text{fat}_{\mathscr{F}}$, and the lower bound of (3) that $\text{fat}_{\mathscr{F}}(\gamma) = 1$ when $\gamma > 1$. The proof of Theorem 1 is finished. □

We now proceed to make the function (1) and (2) of Theorem 1 explicit.

Corollary 1. *Let H denote an inner-product space of dimension n and consider its closed ball $X = B_{R_1} \subset H$ of radius R_1. Let $\mathscr{F}$ denote the class of functions on X defined by $f(x) = \phi \cdot x + c$ with $|\phi|_H \leqslant R_2$ and $c \in \mathbb{R}$. Then*

$$\max\left(\frac{R_1^2 R_2^2}{\gamma^2}, 1\right) \leqslant \text{fat}_{\mathscr{F}}(\gamma) \leqslant \min\left(\frac{R_1^2 R_2^2}{\gamma^2} + \frac{5}{4}, n+1\right).$$

Proof. Again we prove the corollary for $R_1 = R_2 = 1$, and the general theorem then follows from simple scaling arguments.

Equation (2) becomes

$$\text{fat}_{\mathscr{F}}(\gamma) = k, \quad \frac{k+1}{k}\frac{1}{\sqrt{k+2}} < \gamma \leqslant \frac{1}{\sqrt{k-1}}, \quad k \text{ even}, \tag{12}$$

$$\text{fat}_{\mathscr{F}}(\gamma) = k, \quad \frac{1}{\sqrt{k}} < \gamma \leqslant \frac{k}{k-1}\frac{1}{\sqrt{k+1}}, \quad k \text{ odd}. \tag{13}$$

To finish the proof of Corollary 1 we establish

$$\text{fat}_{\mathscr{F}}(\gamma) - \frac{1}{\gamma^2} \in \left(0, \frac{5}{4}\right]. \tag{14}$$

To that end, consider the case where $\text{fat}_{\mathscr{F}}(\gamma) = k$ is even. We prove the sharper statement

$$\text{fat}_{\mathscr{F}}(\gamma) - \frac{1}{\gamma^2} \in \left(\frac{n+1}{(n+2)^2}, 1\right], \quad \text{fat}_{\mathscr{F}}(\gamma) \text{ even}. \tag{15}$$

https://doi.org/10.1017/S0963548303005911 Published online by Cambridge University Press

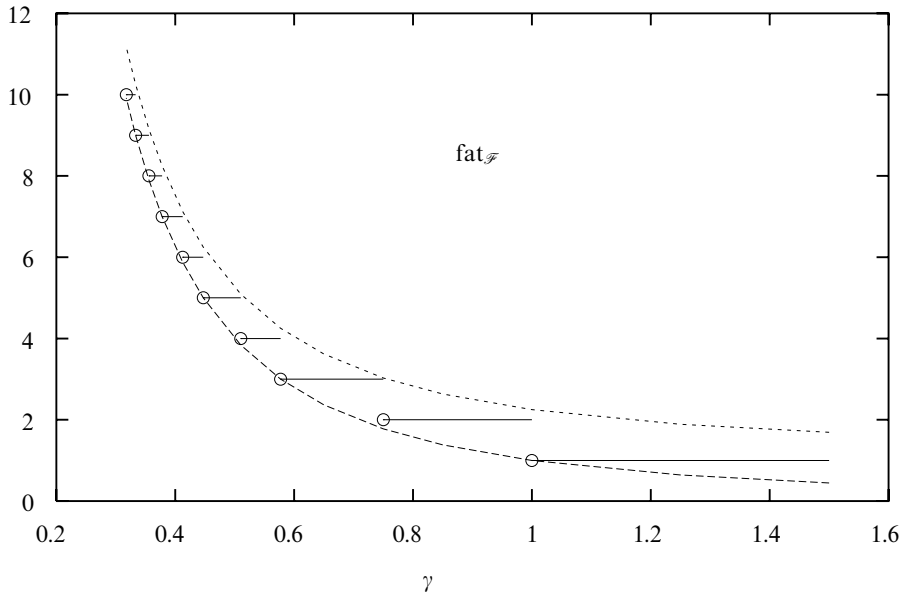


Figure 1.

Equation (12) is equivalent to

$$\frac{k}{(k+1)^2} < k - \frac{1}{\gamma^2} \leqslant 1.$$

Since $\mathrm{fat}_{\mathscr{F}}(\gamma) \leqslant n+1$ and $\frac{k}{(k+1)^2}$ is monotonically decreasing we obtain the claim (15) for finite n. It is not hard to see that the result is also correct when $n = \infty$. Now consider the case where $\mathrm{fat}_{\mathscr{F}}(\gamma) = k$ is odd. Equation (13) is equivalent to

$$0 < k - \frac{1}{\gamma^2} \leqslant 1 + \frac{1}{k} - \frac{1}{k^2}. \tag{16}$$

Since the function $1 + \frac{1}{k} - \frac{1}{k^2}$ has a maximum value $\frac{5}{4}$, we conclude

$$\mathrm{fat}_{\mathscr{F}}(\gamma) \in \frac{1}{\gamma^2} + \left(0, \frac{5}{4}\right], \quad \mathrm{fat}_{\mathscr{F}}(\gamma) \text{ odd}. \tag{17}$$

Note that inequality (16) is sharp in the sense that the choice $\gamma = \frac{k}{k-1}\frac{1}{\sqrt{k+1}}$ achieves the upper bound, and since the function $1 + \frac{1}{k} - \frac{1}{k^2}$ is strictly greater than 1 for $k > 1$ we conclude that the often stated bound

$$\mathrm{fat}_{\mathscr{F}}(\gamma) \leqslant \frac{1}{\gamma^2} + 1 \tag{18}$$

is incorrect.

Combining the even (15) and the odd (17) cases implies the claim (14), and together with (3) completes the proof of Corollary 1. □

https://doi.org/10.1017/S0963548303005911 Published online by Cambridge University Press

In Figure 1 we present a graph of the function $\text{fat}_{\mathcal{F}}$ along with the upper and lower bounds implied by Corollary 1.

References

[1] Alon, N., Ben-David, S., Cesa-Bianchi, N. and Haussler, D. (1997) Scale-sensitive dimensions, uniform convergence, and learnability. *J. Assoc. Comput. Mach.* **44** 615–631.
[2] Bartlett, P. L. and Shawe-Taylor, J. (1999) Generalization performance of support vector machines and other pattern classifiers. In *Advances in Kernel Methods: Support Vector Learning* (B. Schölkopf, C. J. C. Burges and A. J. Smola, eds), MIT Press.
[3] Devroye, L., Györfi, L. and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition*, Springer, New York.
[4] Gurvits, L. (1997) A note on the scale sensitive dimension of linear bounded functionals in banach spaces. In *Proc. Algorithm Learning Theory* (*ALT-97*).
[5] Hush, D. and Scovel, C. (2001) On the VC dimension of bounded margin classifiers. *Machine Learning* **45** 33–44.
[6] Kearns, M. and Schapire, R. E. (1994) Efficient distribution-free learning of probabilistic concepts. *J. Comput. System Sci.* **48** 464–497.
[7] Shawe-Taylor, J., Bartlett, P., Williamson, R. C. and Anthony, M. (1998) Structural risk minimization over data-dependent hierarchies. *IEEE Trans. Inform. Theory* **44** 1926–1940.
[8] Vapnik, V. N. (1998) *Statistical Learning Theory*, Wiley, New York.

https://doi.org/10.1017/S0963548303005911 Published online by Cambridge University Press