On the Expected Number of k-Sets (Abstract)

Imre Bárány[1]
Mathematical Institute
The Hungarian Academy of Sciences
and
Cowles Foundation
Yale University

William Steiger[1,2]
Department of Computer Science
Rutgers University

May 15, 1990

Abstract

Given a set S of n points, a subset X of size k is called a *k-set* if there is a hyperplane Π that separates X from $S - X$. We study $E_d(k;n)$, the expected number of *k-sets* when S is a sample of n points from a distribution F on $\Re^d$. For all the distributions considered, when k is bounded E is asymptotically the expected number of vertices on the convex hull of n random points from F; when k is proportional to n, E is $O(n^{d-1})$.

[1]The authors express gratitude to the NSF DIMACS Center at Rutgers
[2]Research Supported in Part by NSF grant CCR-8902522

1 Introduction and Summary

Let $S = \{x_1, \ldots, x_n\}$ denote n points in R^d. A subset X of size k is called a k-set if there is a hyperplane Π that separates X from $S - X$. Most of the previous work has focused on $e_d(k;n)$, the maximal number of k-sets over all configurations of n points in R^d.

Clearly $\Omega(n^{d-1})$ and $O(n^d)$ provide upper and lower bounds, respectively, for $e_2(k;n)$. Nontrivial were obtained obtained by Lovász [12] for halving sets (n even, $k = n/2$), and later, for general $k \leq n/2$, by Erdős, Lovász, Simmons and Strauss [10]. A simple construction gives a set S with $n\log(k+1)$ k-sets, while a counting argument shows that $e_2(k;n) = O(n\sqrt{k})$. These bounds were rediscovered several times, for example by Edelsbrunner and Welzl [9] but had not been improved until Pach, Steiger, and Szemerédi [13] reduced the bound to $n\sqrt{k}/log^*k$. The papers [1],[11],[15] contain results related to the study of $e_2(k;n)$.

Raimund Seidel (see [8]) extended the Lovász lower bound construction to $d = 3$ and showed that $e_3(k;n) = \Omega(nk\log(k+1))$. The argument may be applied inductively to show $e_d(k;n) = \Omega(nk^{d-2}log(k))$

A non-trivial upper bound for $d = 3$ was recently obtained by Bárány, Füredi, and Lovász [3]. They show that $e_d(n/2;n) = n^{3-\epsilon}$, where $\epsilon > 0$ is some small constant. This, in turn, was improved by Aronov, Chazelle, Edelsbrunner, Guibas, Sharir, and Wenger [2] to $O(n^{8/3}log^{5/3}n)$. For $d > 3$, only the trivial bound is known.

It appears likely that the truth is near the lower bound. Support comes from the fact that in "typical" cases there are relatively few k-sets. We study $E_d(k;n)$, the expected number of k-sets when S is a sample of n random points from a distribution F on R^d. We deal with the cases where F is a spherically symmetric distribution or the uniform distribution in a convex polhedron. In each case, for bounded k, E grows like the expected number of vertices on the convex hull of n points with distribution F, and when k is proportional to n, E grows like n^{d-1}. We mention one related paper by Sharir [14]. Let N points in the plane be given. Using a random sampling argument, he shows that suitable random subsets of size n are expected to have at most $O(n^{6/5+\delta})$ halving sets, a statement free of any assumptions concerning distributions.

Despite the simplicity of the methods we use to derive these results, the statements seem to be the first known facts about $E_d(k;n)$. Combined with the fact that k-sets have applications in computational geometry and machine learning, we feel that this work might be useful and interesting.

2 Results

If F is a distribution function on R^d, $F(S)$ denotes the probability F assigns to the Borel set $S \subseteq R^d$. We have a sample of size $n+d$ from F. Given d specific points $x_1, \ldots, x_d$,

they determine a hyperplane H. Write $Z(x_1,\ldots,x_d) = Prob_F(h)$, for the random variable measuring the F-probability of h, the smaller of the two open halfspaces determined by H, and let $G(t)$ denote its distribution, $t \leq 1/2$.

The key relation describes the probability P, that $x_1,\ldots,x_d$ determine a k-set. We have

$$P = \binom{n}{k} \int_0^{1/2} [t^k(1-t)^{n-k} + (1-t)^k t^{n-k}] dG(t), \tag{1}$$

dG denoting Stieltjes integration. The integrand describes the probability that k points are in h and $n-k$ in its complement, or vice-versa, times the probability that h has has probability content t. It then follows that

$$E_d(k; n+d) = \binom{n+d}{d} P.$$

All statements depend on the analysis of G and the integrand in (1). For example if G(t) = 2t, the integral in (1) is $Beta(k+1, n-k+1)$ and P is $(n+1)^{-1}$. Suppose first that $d = 2$.

Lemma 1 *If F is circularly symmetric then*

$$E_2(k; n+2) \leq n+2.$$

The proof is a calculation using the fact that $G(t+u) - G(t) \leq 2u$ and the identity

$$\int_0^1 t^k(1-t)^{n-k} dt = \left[(n+1)\binom{n}{k}\right]^{-1}.$$

Here is a sketch. Suppose F is circularly symmetric about the origin. For each $t \in (0, 1/2)$ there is a disk C_t centered at the origin with radius $u(t)$ determined by the property that each tangent defines a halfspace which has probability content t.

Let $A(t)$ denote the complement of C_t. For each point $x \in A(t)$, there are two tangents to C_t, $\tau_1(x)$ and $\tau_2(x)$, each defining a halfspace of probability t. Let $W_t(x)$ denote the symmetric difference of these halfspaces; these are the points y such that the segment xy defines a halfspace of probability $\leq t$. Then

$$G(t) = \int_{x \in A(t)} Prob_F(W_t(x)) dF(x). \tag{2}$$

To estimate $G(t+u) - G(t)$, $u > 0$, use (2) to see that

$$\begin{aligned} G(t+u) - G(t) &= \int_{x \in A(t)} Prob_F(W_{t+u}(x) - W_t(x)) dF(x) \\ &+ \int_{x \in A(t+u) - A(t)} Prob_F(W_{t+u}) dF(x). \end{aligned} \tag{3}$$

The statement of the lemma easily follows by bounding each integral in (3) by u. In the first case the integrand is at most u. In the second, the range of integration has probability less than u. If F is absolutely continuous w.r.t. Lebesgue measure then $E_2(k;n)$ is bounded below by cn for some constant $c \in (0,1)$. In addition this kind of calculation can be made in R^d to reveal that when F is spherically symmetric, $E_d(k;n) = O(n^{d-1})$.

The integrand in (1) is strongly concentrated around $t = k/n$. This enables us to make more exact statements when there is more detailed information about G, for example when F is the uniform distribution in some bounded set. In one such case we can prove

Lemma 2 *If F is the uniform distribution in the unit circle, then*

$$E_2(k;n) = \Theta(k^{2/3}n^{1/3})$$

There is an analogous statement for the ball in R^d. Note that for fixed k, this expression is asymptotic to $n^{1/3}$, the expected number of hull vertices for n random points sampled from F; for k proportional to n it agrees with Lemma 1.

In a similar way,

Lemma 3 *If F is the uniform distribution on the unit square*

$$E_2(k;n) = \Theta(k \log n / \log k)$$

There is an analogous statement for uniform distributions in other convex polygons and we can generalize to higher dimensions. All the results are established by exploiting the uniformity of F and the concentration of the integrand of (1).

Finally we note that it is an interesting question as to whether there are distributions F for which $E_d(k;n)$ is of a strictly higher order than n^{d-1}. We have not been able to find one.

References

[1] N. Alon and E. Győri. The number of small semispaces of a finite set of points. *J. Combin. Theory A*, 41:154–157, 1986.

[2] B. Aronov, B. Chazelle, H. Edelsbrunner, L. Guibas, M. Sharir, and R. Wenger. Points and triangles in the plane and halving planes in space. Preprint, 1990.

[3] I. Bárány, Z. Füredi, and L. Lovász. On the number of halving planes in R^3. *Proc. Fifth ACM Symposium on Computational Geometry*, pages 140-144, 1989.

[4] B. Chazelle and F. Preperata. Halfspace range search:an algorithmic application of k-sets. *Discrete and Comp. Geom.*, 1:83–93, 1986.

[5] K. Clarkson. New applications of random sampling in computational geometry. *Discrete and Comput. Geom.*, 2:195–222, 1987.

[6] K. Clarkson and P. Shor. Applications of random sampling in computational geometry II. *Discrete and Comput. Geom.*, 2:387–421, 1989.

[7] R. Cole, M. Sharir, and C. Yap. On k-hulls and related problems. *SIAM J. Computing*, 16:61–77, 1987.

[8] H. Edelsbrunner. *Algorithms in Combinatorial Geometry.* Springer-Verlag, Berlin, 1987.

[9] H. Edelsbrunner and E. Welzl. On the number of line separations of a finite set in the plane. *J. Combin. Theory A*, 38:15–29, 1985.

[10] P. Erdös, L. Lovász, A. Simmons, and E. Strauss. Dissection graphs of planar point sets. In J. Srivastava and et.al., editors, *A Survey of Combinatorial Theory*, pages 139–149. North-Holland, Amsterdam, 1973.

[11] J.E. Goodman and R. Pollack. On the number of k-sets of a set of n points in the plane. *J. Combin. Theory A*, 36:101–104, 1984.

[12] L. Lovász. On the number of halving lines. *Ann. Univ. Sci. Budapest, Eötvös, Sect. Math.*, 14:107–108, 1971.

[13] J. Pach, W. Steiger, and E. Szemerédi. An Upper Bound on the Number of Planar k-Sets. *Discrete and Comp. Geom. (to appear, 1990).*

[14] M. Sharir. Randomized Analysis of k-Sets.

[15] E. Welzl. More on k-sets of finite sets in the plane. *Discrete and Comput. Geom.*, 1:95–100, 1986.