(revised by the authors)

Lower Bounds on the Vapnik-Chervonenkis Dimension of Convex Polytope Classifiers

Gábor Takács, Béla Pataki
Budapest University of Technology and Economics
Department of Measurement and Information Systems
Magyar tudósok körútja 2., Budapest, H-1117, Hungary
gtakacs@mit.bme.hu, pataki@mit.bme.hu

***Abstract*— In statistical learning theory, the Vapnik-Chervonenkis (VC) dimension is an important combinatorial property of classifier families. In this paper we examine the case of convex polytope classification, i.e. when the separation of the two classes is done by a convex surface consisting of linear segments. We collect the known facts about the VC dimension of convex polytope classifiers with n facets in $\mathcal{R}^d$ and present two new lower bounds (one for the general case and one for the special case $d = 4$).**

I. Introduction

In this article we will consider the problem of two-class classification. In this setting a *classifier* is an $\mathcal{R}^d \mapsto \{+1, -1\}$ mapping, where the input vector can be referred as the *observation* and the assigned value can be called the *class label*. Clearly, every classifier g partitions the space $\mathcal{R}^d$ into two distinct regions. It is an important special case, when the separation is done by a linear surface, i.e. the function g can be written as

$$g(\mathbf{x}) = \mathrm{sgn}(\mathbf{w}^T\mathbf{x} + b),$$

where

$$\mathrm{sgn}(y) = \begin{cases} +1 & \text{if } y \geq 0, \\ -1 & \text{if } y < 0. \end{cases}$$

Such classifiers are said to be *linear*. The set of all d-dimensional linear classifiers is denoted by LIN(d).

We get a richer function class, if the label is decided by combining the outputs of n linear classifiers:

$$g(\mathbf{x}) = f(l_1(\mathbf{x}), l_2(\mathbf{x}), \ldots, l_n(\mathbf{x})),$$

$$l_i(\mathbf{x}) = \mathrm{sgn}(\mathbf{w}_i^T\mathbf{x} + b_i), \quad i = 1, 2, \ldots, n,$$

where f is an arbitrary $\{+1, -1\}^n \mapsto \{+1, -1\}$ mapping, called the *composition function*. Classifiers that can be expressed in this form are called *polytope classifiers*. The set of all d-dimensional polytope classifiers using n linear decisions is denoted by POL(d, n). The function class POL(d, n) is quite extensive: it contains e.g. linear classifiers, decision trees and nearest neighbor classifiers. We obtain interesting subsets of POL(d, n), if we restrict the set of available composition functions to minimum and maximum taking:

$$\begin{aligned} \mathrm{MIN}(d, n) &= \{g : g \in \mathrm{POL}(d, n), f = \min\}, \\ \mathrm{MAX}(d, n) &= \{g : g \in \mathrm{POL}(d, n), f = \max\}, \\ \mathrm{MINMAX}(d, n) &= \{g : g \in \mathrm{POL}(d, n), f \in \{\min, \max\}\}. \end{aligned}$$

It is true for all of the above cases that the decision boundary is the surface of a d-dimensional convex polytope. The label of the inner (convex) region is always +1 at MIN(d, n), always -1 at MAX(d, n) and unrestricted in the case of MINMAX(d, n). The members of MINMAX(d, n) are called *convex polytope classifiers*. Note that the decision boundary belongs to the inner region in the case of MIN(d, n) and to the outer region in the case of MAX(d, n).

In statistical learning theory [1], the Vapnik-Chervonenkis (VC) dimension is an important combinatorial property of classifier families. We say that a set of classifiers $\mathcal{G}$ *shatters* a finite set of observations, if the observations can be arbitrarily labeled by the members of $\mathcal{G}$. The kth *shatter coefficient* of $\mathcal{G}$ (denoted by $s(\mathcal{G}, k)$) is the maximum number of different labelings that can be produced by the members of $\mathcal{G}$ over any k observations. The Vapnik-Chervonenkis (VC) dimension of $\mathcal{G}$ (denoted by $h(\mathcal{G})$) is the maximum number of observations that can be shattered by $\mathcal{G}$. Formally:

$$h(\mathcal{G}) = \max k : s(\mathcal{G}, k) = 2^k.$$

If this maximum does not exist, then the VC dimension is ∞.

The concept of VC dimension is very useful in the field of classification, because it appears in distribution-free error bounds. Given a classifier g, there is no general connection between its error probability $R(g)$ and its error rate $R_m(g)$ measured on the m-element training set. However if we know a priori that $g \in \mathcal{G}$ and $h(\mathcal{G}) < \infty$, then the following inequality holds with probability $1 - \delta$ [1]:

$$R(g) \leq R_m(g) + \sqrt{8\frac{\mathrm{h}(\mathcal{G})\ln(2em/\mathrm{h}(\mathcal{G})) + \ln(2/\delta)}{m}}.$$

Convex polytope classification is not a new invention of course. Beside the different algorithms (e.g. [2], [3], [4], [5]), numerous results can be found in the literature about the combinatorial properties of MIN(d, n) and MAX(d, n) (see e.g. [2], [6], [7], [8]). The function class MINMAX(d, n) is an undiscovered area to the best of the authors' knowledge. However, this class stays closer to practice than MIN(d, n) or MAX(d, n), because the label of the inner (convex) region is not fixed. The goal of this paper is to collect the known facts about the VC dimension of MIN(d, n), MAX(d, n) and MINMAX(d, n), while presenting two new lower bounds (one for the general case and one for the special case $d = 4$).

II. Known facts

Clearly, the VC dimension of MIN(d, n) and MAX(d, n) are equal, therefore we will state the theorems only for MIN(d, n) and MINMAX(d, n). Let us start with the simplest special cases: $n = 1$ and $d = 1$! If $n = 1$, then clearly, MIN(d, n) = MINMAX(d, n) = LIN(d). It is known since long ago that $h(\mathrm{LIN}(d)) = d + 1$ [6]. If $d = 1$ and $n > 1$, then trivially $h(\mathrm{MIN}(d, n)) = 2$ and $h(\mathrm{MINMAX}(d, n)) = 3$.

Now let us proceed with the case $d = 2$! Determining the VC dimension of MIN(2, n) is easy:

Theorem 1: If $n > 1$, then $h(\mathrm{MIN}(2, n)) = 2n + 1$.

Proof. If we place $2n+1$ points into the vertices of a regular $(2n+1)$-gon, then every labeling can be produced by n lines. This implies $h(\mathrm{MIN}(2, n)) \geq 2n + 1$. For proving the upper bound, consider an arbitrary point set $\mathcal{P}$ of size $(2n + 2)$. If one of the points is located inside the convex hull of the others, then $\mathcal{P}$ cannot be shattered by MIN(2, n), because it is impossible to classify the inside point as -1 and the outside points as +1. If the convex hull of $\mathcal{P}$ is a convex $(2n+2)$-gon, then $\mathcal{P}$ cannot be shattered by MIN(2, n) again, because it is impossible to realize the alternating labeling with n lines. ■

With considerably more effort it can be proved too that $h(\mathrm{MINMAX}(2, n)) = 2n + 2$, if $n > 1$ [9]. In any other cases, the exact value of the VC dimension is unknown.

For the case $d = 3$ the best bounds originate from Dobkin and Gunopulos. In [2] they showed that $h(\mathrm{MIN}(3, n)) \leq 4n$, and presented a construction that proves $h(\mathrm{MIN}(3, 4)) \geq 14$. With a straightforward extension of their construction it can also be shown that $h(\mathrm{MIN}(3, n) \geq 3n + 2$ and $h(\mathrm{MINMAX}(3, n)) \geq 3n + 3$, if $n \geq 4$.

In the general case we can easily get an upper bound with combinatorial tools:

Theorem 2: $h(\mathrm{MIN}(d, n)) \leq 2(d + 1)n \ln(3n)$.

Proof. We will exploit the fact that MIN(d, n) classifiers consist of simple components that are combined in a simple way. Denote the kth shatter coefficient of MIN(d, n) by S_k and that of LIN(d, n) by s_k. Given k points, the component classifiers are able to produce at most s_k different labelings. This implies that

$$S_k \leq (s_k)^n.$$

By Sauer's lemma [10] we know that

$$s_k \leq \sum_{i=0}^{d+1} \binom{k}{i}.$$

The sum of binomial coefficients can be bounded from above as follows:

$$\sum_{i=0}^{d+1} \binom{k}{i} < \left(\frac{k}{d+1}\right)^{d+1} \sum_{i=0}^{d+1} \binom{k}{i} \left(\frac{d+1}{k}\right)^{i} <$$
$$\left(\frac{k}{d+1}\right)^{d+1} \left(1 + \frac{d+1}{k}\right)^{k} < \left(\frac{ek}{d+1}\right)^{d+1}.$$

Putting the inequalities together yields

$$S_k \leq (s_k)^n < \left(\frac{ek}{d+1}\right)^{(d+1)n}.$$

If $k = 2(d + 1)n \log 3n$, then it can be easily verified that

$$\left(\frac{ek}{d+1}\right)^{(d+1)n} < 2^k.$$

■

The straightforward way of obtaining an upper bound for $h(\mathrm{MINMAX}(d, n))$ is applying Assouad's lemma [11]. Assouad's lemma states that for any function classes $\mathcal{F}$ s $\mathcal{G}$ with finite VC dimension, $h(\mathcal{F} \cup \mathcal{G}) \leq h(\mathcal{F}) + h(\mathcal{G}) + 1$. In the case of MINMAX($d$, n) this means that $h(\mathrm{MINMAX}(d, n)) \leq h(\mathrm{MIN}(d, n)) + h(\mathrm{MAX}(d, n)) + 1 = 4(d + 1)n \ln(3n) + 1$.

III. New Lower Bounds

At first we show two easily obtainable lower bounds for $h(\mathrm{MIN}(d, n))$ and $h(\mathrm{MINMAX}(d, n))$. They are so simple that we will call them the trivial lower bounds. (However, to the best of our knowledge they have not been published yet.)

Theorem 3 (sphere slicing): $h(\mathrm{MIN}(d, n)) \geq dn$.

Proof.[1] We have to arrange dn points such that they can be shattered by MIN(d, n). At first cut n distinct slices from a d-dimensional hypersphere with n hyperplanes. Denote the ith hyperplane with $\mathcal{H}_i$ and denote the intersection of the hypersphere surface and the ith hyperplane with $\mathcal{C}_i$.

Now define the point set $\mathcal{P}$ by assigning d different points in each $\mathcal{C}_i$. Points belonging to the same $\mathcal{C}_i$ can be shattered by one hyperplane, moreover, it can be required too, that the label of all the other points have to be +1. (The ith hyperplane of the MIN(d, n) classifier is obtained by an infinitesimal perturbation of $\mathcal{H}_i$.) Because of the requirement, the ith shattering hyperplane influences only the the labeling of the ith group, therefore $\mathcal{P}$ can be arbitrarily labeled by the members of MIN(d, n). ■

Theorem 4: $h(\mathrm{MINMAX}(d, n)) \geq dn + 1$.

Proof. Arrange the first dn points like in the proof of Theorem 3, but now put an additional point into the center of the sphere. Because of Theorem 3, the first dn points can be shattered by both MIN(d, n) and MAX(d, n). In the first case the label of the central point is always +1, in the second case it is always -1. This means, that the $dn+1$ points can be arbitrarily labeled by the members of MINMAX(d, n). (The label of the central point determines whether to use a MIN(d, n) or a MAX(d, n) classifier.) ■

The following theorem is a new result that improves the trivial lower bounds by one:

Theorem 5: If $n \geq 2$, then $h(\mathrm{MIN}(d, n)) \geq dn + 1$ and $h(\mathrm{MINMAX}(d, n)) \geq dn + 2$.

Proof. We begin with the proof of the first statement in the special case $d = 3$. Place the first 5 points on the plane xy, into the vertices of an origin-centered regular pentagon. On the plane xy these points can be shattered by 2 lines, moreover, the followings can be satisfied too:

- The lines have to be parallel.
- The lines cannot be axis-parallel.

[1] The proof originates from István Pilászy, a colleague of the authors.

- For any labeling, the ball of radius $\varepsilon_0 > 0$ around the origin must belong to the inner region.

Place the next 2 points into $[1 \quad 0 \quad 1]$ and $[-1 \quad 0 \quad 1]$. Now we show that the first 7 points can be arbitrarily labeled by the members of MIN(3, 2), even if we require that an ε-ball around the origin have to belong to the inner region.

Consider an arbitrary labeling $y_1, y_2, \ldots, y_7$ that we want to realize by a member of MIN(3, 2). Recall that a MIN(3, 2) classifier can be given by 8 parameters: $(w_{11}, w_{12}, w_{13}, b_1)$ and $(w_{21}, w_{22}, w_{32}, b_2)$. If the input is one of the first 5 points, then the answer of the classifier is fully determined by the parameters (w_{11}, w_{12}, b_1) and (w_{21}, w_{22}, b_2). Adjust these parameters such that the classifier labels the first 5 points in the desired way. This can be done, moreover the followings can be satisfied too $(i = 1, 2)$:

$$w_{2i} = -w_{1i} \neq 0,$$
$$\forall \delta_1^2 + \delta_2^2 \leq \varepsilon_0^2 : \min_i \{\delta_1 w_{i1} + \delta_2 w_{i2} + b_i\} \geq 0.$$

Based on y_6 and y_7 there are 4 possible cases $(i = 1, 2)$:

- If $y_6 = y_7 = +1$, then the desired labeling can be obtained by $w_{i3} \gg 0$.
- If $y_6 = y_7 = -1$, then the desired labeling can be obtained by $w_{i3} \ll 0$.
- If $y_6 = +1, y_7 = -1$, then the desired labeling can be obtained by $w_{i3} = -w_{i1} - b_i + \mu$.
- If $y_6 = -1, y_7 = +1$, then the desired labeling can be obtained by $w_{i3} = +w_{i1} - b_i + \mu$.

If we choose a sufficiently small ε, then the ε-ball around the origin belongs to the inner region in all of the 4 cases.

Now define a sphere surface that encloses the 7 points arranged so far and passes through the point $[0 \quad 0 \quad -\varepsilon/2]$. On this surface there exist a segment, that belongs to the inner region at each labeling of the first 7 points. Let us place the remaining $3(n-2)$ points onto this subsurface in 3-element groups with the sphere slicing method! Each group can be shattered by one plane, moreover it can be assured, that the planes does not affect the labeling of the other groups and the first 7 points. Therefore the $7 + 3(n-2) = 3n + 1$ points can be shattered by MIN(3, n).

The proof of the d-dimensional case is completely analogous with the 3-dimensional one. Now the first 5 points are placed onto the plane of the first 2 coordinates. The next $2(d-2)$ points are arranged in the following way:

$$\begin{aligned}
x_6 &= [+1 \quad 0 \quad 1 \quad 0 \quad 0 \quad \ldots \quad 0] \\
x_7 &= [-1 \quad 0 \quad 1 \quad 0 \quad 0 \quad \ldots \quad 0] \\
x_6 &= [+1 \quad 0 \quad 0 \quad 1 \quad 0 \quad \ldots \quad 0] \\
x_9 &= [-1 \quad 0 \quad 0 \quad 1 \quad 0 \quad \ldots \quad 0] \\
&\vdots \\
x_{2d} &= [+1 \quad 0 \quad 0 \quad 0 \quad \ldots \quad 0 \quad 1] \\
x_{2d+1} &= [-1 \quad 0 \quad 0 \quad 0 \quad \ldots \quad 0 \quad 1].
\end{aligned}$$

It can be shown (exactly the same way as in the case $d = 3$) that the first $2d + 1$ points can be shattered by MIN(d, 2), moreover it can be required too that an ε-ball around the origin belongs to the inner region at each labeling. Then a hypersphere surface is defined that encloses the first $2d + 1$ points and passes through the point $[0 \quad 0 \quad -\varepsilon/2 \quad \ldots \quad -\varepsilon/2]$. There exist a segment on this surface that belongs to the inner region at each labeling. The remaining $d(n-2)$ points are placed onto this subsurface with the sphere slicing method.

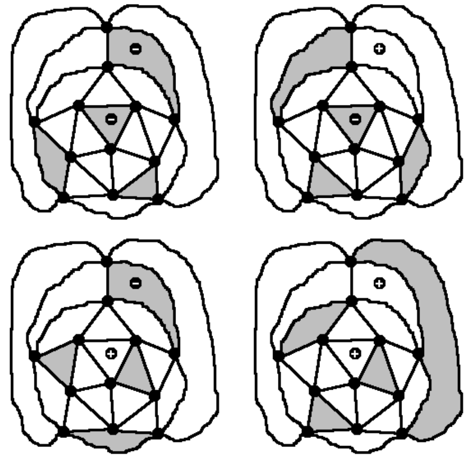

Fig. 1. This figure shows the planar graph of the icosahedron and illustrates how to choose the independent faces based on the label of the extra points. Note that there exist faces that are never chosen.

The statement on MINMAX(d, n) can be proved with the same construction with an additional point in the origin. ■

The bound $h(\mathrm{MIN}(d, n)) \leq dn + 1$ is not tight. In $\mathcal{R}^3$ a better bound can be obtained with a tricky arrangement [2]:

Theorem 6 (Dobkin, Gunopulos): $h(\mathrm{MIN}(3, 4)) \geq 14$.

Proof. Let us place the first 12 points into the vertices of a regular icosahedron! Clearly, these points can be shattered by MIN(3, 4). (The 4 planes are obtained by infinitesimal perturbation of 4 independent faces.) Now place 2 extra points outside the the icosahedron but close to its surface, above 2 faces at face distance 3 from each other. Surprisingly, this larger point set can still be arbitrarily labeled by the members of MIN(3, 4). This can be easily verified if we draw the planar graph of the icosahedron. The trick is that the 4 independent faces are chosen based on the label of the 2 extra points (see Fig. 1). ■

This result can easily be extended to the case $n > 4$.

Theorem 7: If $n \geq 4$, then $h(\mathrm{MIN}(3, n)) \geq 3n + 2$ and $h(\mathrm{MINMAX}(3, n)) \geq 3n + 3$.

Proof. Place the first 14 points as in the proof of Theorem 6. From the proof of Theorem 6 we know that these points can be shattered by MIN(3, 4) and there exist a face of the icosahedron that is never selected. This means that we can define a sphere surface that encloses the first 14 points and has a segment that is always in the inner region. If we place the remaining $3(n-4)$ points on this subsurface with the sphere slicing method, then this point set of size $14 + 3(n-4) = 3n + 2$ can be shattered by MIN(3, n).

The second statement of the theorem can be proved with almost the same construction. The difference is that if we can use MAX(3, n) classifiers too, then we can put an additional point into the center of the icosahedron. ■

With a more complicated version of the icosahedron trick it is possible to improve the trivial lower bounds in $\mathcal{R}^4$ by 4. The following theorem is our second new result:

Theorem 8: If $n \geq 30$, then $h(\mathrm{MIN}(4, n)) \geq 4n + 4$ and $h(\mathrm{MINMAX}(4, n)) \geq 4n + 5$.

Proof. First of all we recall from geometry that the *600-cell* is the finite regular four-dimensional polytope with Schläfli symbol $\{3, 3, 5\}$ [12]. It is also called *hypericosahedron*, because it can be regarded as the 4-dimensional analog of the icosahedron. The 600-cell is composed of 600 tetrahedra, with 5 to an edge. It has 1200 triangular faces, 720 edges and 120 vertices.

The vertices of an origin-centered 600-cell with edges of length $1/\phi$ (where $\phi = (1 + \sqrt{5})/2$ is the golden ratio) can be given as follows:

- 16 vertices of the form $[\pm 0.5 \quad \pm 0.5 \quad \pm 0.5 \quad \pm 0.5]$.
- The 8 possible permutations of $[\pm 1 \quad 0 \quad 0 \quad 0]$.
- 96 vertices, obtained from the even permutations of $[\pm 0.5 \quad \pm 0.5\phi \quad \pm 0.5/\phi \quad 0]$.

The topological structure of the 600-cell is a system of subsets over the 120 vertices that gives which k vertices form a k-facet ($k = 2, 3, 4$; 2-facets are called edges, 3-facets are called faces and 4-facets are called cells). The topological structure of the 600-cell can be computed easily: At first generate the coordinates of the 120 vertices according to the previous scheme. Then identify the k-facets by examining every possible k vertices an checking whether they are at distance $1/\phi$ from each other not.

We say that 2 cells are adjacent, if they have at least 1 common vertex and 2 cells are independent, if they have no common vertices. The cell adjacency graph of the 600-cell can be easily obtained from its topological structure. Remember that in the 3-dimensional case we tried to cover the vertices of the icosahedron with independent facets in many different ways. Now we want to cover the vertices of the 600-cell with independent cells in many possible ways. This means that we want to find many independent points in the cell adjacency graph.

With brute force computation we could find 1920 different coverings.[2] These coverings can be represented by a 1920-by-600 binary matrix C, whose element c_{ij} is 1 if and only if the ith covering contains the j cell. We will refer to C as the covering matrix.

Now we turn back to the proof of the statement $h(\mathrm{MIN}(4, n)) \geq 4n + y$, if $n \geq 30$. Let us place the first 120 points into the vertices of a 600-cell! These points can be shattered by MIN(4, 30), because it is possible to cover the 120 points with 30 independent cells. (The different labelings are obtained by infinitesimally perturbing the 30 independent cells.)

[2]The topological structure of the 600-cell, the cell adjacency graph and the coverings can be seen at the website http://www.mit.bme.hu/~gtakacs/600cell.html. It is quite sweaty to verify these results by hand but it is easy to write a program that performs this.

If there exist an l-column submatrix in the covering matrix that has 2^l different rows, then l extra points can be placed into this arrangement such that the point set can still be shattered by MIN(4, 30). We wrote a simple program that counts the number of different rows in every possible l-column submatrix. The largest l for that an appropriate subset of columns could be selected was $l = 4$. This means that it is possible to arrange 120 + 4 = 124 points in $\mathcal{R}^4$ such that they can be shattered by MIN(4, 30).

Like the previous theorems we can define a 4-dimensional sphere that encloses the first 124 points and contains a surface segment that belongs to the inner region at each labeling of the first 124 points. If the remaining $4(n - 30)$ points are placed onto this surface segment with the sphere slicing method, then the $dn + 4$ can be arbitrarily labeled by MIN(4, n) assuming that $n \geq 30$. In the case of MINMAX(4, n) the construction is almost the same, but an additional point can be placed into the origin too. ■

IV. Conclusion

The Vapnik-Chervonenkis dimension of convex polytope classifiers was investigated in this paper. We defined 3 function classes MIN(d, n), MAX(d, n) and MINMAX(d, n), collected the known facts about the VC dimension of these classes and presented two novel lower bounds.

It was shown in this article that the trivial lower bound for $h(\mathrm{MIN}(d, n))$ is dn, and an easily obtainable upper bound is $O(dn \log(n))$. It was mentioned too that the $\log(n)$ factor can be eliminated, if $d < 4$. It is an interesting open problem whether the $\log(n)$ factor can be eliminated or not in the general case. The new results presented in this paper may give some ideas for those who want to answer this question.

References

[1] O. Bousquet, S. Boucheron, G. Lugosi, "Introduction to statistical learning theory", *Lecture Notes in Artificial Intelligence*, vol. 3176, pp. 169–207, 1999.

[2] D. P. Dobkin, D. Gunopulos, "Concept learning with geometric hypotheses", *Proc. 8th Annual Conference on Computational Learning Theory*, pp. 329–336, ACM Press, New York, 1995.

[3] P. Fischer, "More or less efficient agnostic learning of convex polygons", *Proc. 8th Annual Conference on Computational Learning Theory*, pp. 228–236, ACM Press, New York, 1995.

[4] S. Kwek, L. Pitt, "PAC learning intersections of halfspaces with membership quieries" *Algorithmica*, vol. 22, pp. 53–75, 1998.

[5] M. Elad, Y. Hel-Or, R. Keshet, "Rejection based classifier for face detection", *Pattern Recognition Letters*, vol. 23, pp. 1459–1471, 2002.

[6] D. Haussler, E. Welzl, "Epsilon nets and simplex range queries", *Discrete Computational Geometry*, vol. 2, pp. 127–151, 1987.

[7] A. Blumer, A. Ehrenfeucht, D. Haussler, M. Warmuth, "Learnability and the Vapnik-Chervonenkis dimension", *Journal of the ACM*, vol. 36, pp. 929–965, 1989.

[8] L. Devroye, L. Györfi, G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York, 1996.

[9] G. Takács, "The Vapnik-Chervonenkis dimension of convex n-gon classifiers", Submitted to: *Hungarian Electronic Journal of Sciences*, 2007.

[10] N. Sauer, "On the density of families of sets", *Journal of Combinatorial Theory (A)*, vol. 13, pp. 145–147, 1972.

[11] P. Assouad, "Densité et dimension", *Annales de l'Institut Fourier*, vol. 33, pp. 233–282, 1983.

[12] H. Coxeter, *Introduction to Geometry, 2nd ed.*, Wiley, New York, 1969.