«LE MATEMATICHE»
Vol. XLIII (1988) - Fasc. I, pp. 195-224

AUTOMATED THEOREM PROVING IN ELEMENTARY GEOMETRY

ALFREDO FERRO (Catania) - GIOVANNI GALLO (New York)

A survey of the main procedures for automatic theorem proving in geometry is presented.

1. Introduction.

One of the most basic tasks in Artificial Intelligence is the possibility of automatically deciding if a given conclusion follows necessarily from some premises. This field of research is called Theorem Proving and has many applications such as Program Correctness, Logic Programming, Expert Systems, Real-Time Systems, Problem Solving etc... Automatic Theorem Proving has received a great impulse in the last 25 years after the pioneering work of H. Wang, J. Gilmore, M. Davis, H. Putman and J.R. Robinson who invented the mostly used general purpose theorem proving procedure: the Resolution Principle. This method has been successfully applied to several computer science areas [29].

Non-Resolution theorem proving procedures have been also designed and experimented for particular mathematical theories.

However the first significant achievement in this direction is due to Wu Wen Tsun [31] who, in 1978, presented a procedure to automatically prove statements of Elementary Geometry. Later on a student of his S.C. Chou in his Ph.D. dissertation at the University of Texas gives an impressive collection of about 350 elementary geometry theorems which he was able to prove by Wu's method. This evident success stimulated a strong interest in this field and several new approaches to geometry theorem proving have been developed in the last five years.

In this paper we survey all of these algorithms. In the first section we describe an interactive theorem prover which was developed by Gelernter in 1959. This method is based on the classical euclidean approach to elementary geometry in the style of modern Prolog-like theorem provers.

In the remaining sections we describe algebraic theorem provers all based on the Cartesian Method consisting of transforming a geometric statement into an equivalent algebraic one via a system of coordinates. This algebraic counterpart consists of showing that a given polynomial (the conclusion) belongs to the radical of the ideal generated by a finite number of polynomials (hypotheses). This kind of approach is used in the methods of Wu, Kapur-Kutzler-Stifter and Carrà-Gallo. In Wu's method an algorithm of J.F. Ritt is used to solve the above problem. Ritt's algorithm [25] was originally given in connection with the calculation of a so called *characteristic set*: a basic step in a constructive approach to the problem of factorizing algebraic varieties.

In Kapur-Kutzler-Stifter method the membership problem for the radical of an ideal of polynomials is solved by Grobner Bases calculation using Buchberger's Algorithm. Finally in Carrà-Gallo method the geometric theorem is proven by computing the dimensions of the hypothesis variety and of the intersection of this variety with the thesis variety and comparing them.

On the other hand it can be proven that any geometric theorem can be reduced to the verification of a single polynomial identity. This identity is generally given in an implicit form and can be verified in several ways. One can try to show that the given polynomial has a very large root as in the *proving by example* method of J. Hong. Another more immediate way is to show that the given polynomial has too many roots [26] [33].

Finally we observe that when these methods give rise to computationally unfeasible computations (and this we expect to happen in the most interesting cases) then a probabilistic approach can be used in order to get a «certified» conjecture. Results of J.T. Schwartz in this direction are described in the final section of the paper.

1. The Gelernter's prover.

The first program able to produce a formal proof of statements of Elementary Geometry was developed in 1959 by H. Gelernter at the IBM Research Center in New York ([13], [14], [15]). It was one of the first large program written in FORTRAN (and one of the very few programs of A.I. implemented in this language). The program used several different ideas (rewriting rules, reduction strategies, analysis of symmetries, study of examples, etc...) to obtain proofs of theorems of elementary geometry. These ideas were largely used in almost all the subsequent provers.

The program had access to a database of axioms, and of proved theorems to be used as a *reduction operator* on the theorems under consideration. For example to prove that two segments are congruent the prover could try to prove that the two segments are corresponding edges of two congruent triangles.

This example shows that the program does not always substitute a complex property by a simpler one, but sometimes it goes into the opposite direction. It was than essential for detecting loops of deductions, to give to the prover a suitable *strategy*.

For the sake of efficiency it was also crucial to recognize, at the very early stages of the computation, the possible symmetries of the problem under consideration and to translate them into syntactic conditions. This has a significant impact on the size of the problems that can be realistically studied by the prover.

Sometimes the simple analysis of the given geometrical entities was not enough to produce the requested proof. Some further constructions had to be carry out. Gelernter's program included such a possibility in a sort of interactive dialogue between the prover and the user.

The general strategy used by the prover was to start from the goal statement by generating several subgoals, and then recursively continue in the same way. The basis of the recursion was achieved by proving the subgoals using the collection of elementary facts stored in the database or considering one or more examples, that were given together with the goal. If later this approach did not reach a conclusion, or a loop in the deduction was detected then an appropriate action of the user ought to be specified.

The analysis of the example performed by the prover was, a fortiori, approximated, and in this part of the execution the features of the newly born FORTRAN played an essential role. Moreover to avoid the loss of generality a sufficiently large collection of examples was supposed to be available to the prover. However this expedient was not sufficient to avoid the introduction of some arbitrariness in the proofs.

The importance of the method that has been briefly described above is principally historical. Many of the ideas and tricks used in this prototype have been refined and improved in later provers and have been applied also in different settings. Indeed it is not difficult to recognize in the primitive reduction strategy outlined above an ancestor of the techniques used in the more recent PROLOG-like provers.

Many of the problems that this prover had to face were overcome by the so called algebraic provers that will be presented in the following sections. For example the problem of analyzing the symmetries is solved in the algebraic provers by a suitable choice of the system of coordinates. Algebraic provers, however, have their own drawbacks too and it seems that an integration of the algebraic methods and the logical techniques is the most promising direction for future geometry theorem provers.

2. General aspects of an algebraic prover.

In this section the main ideas and problems of provers based on algebraic techniques are presented. In the next section a more detailed account of some specific methods will be given.

The very first step for any algebraic prover is a preliminary translation of a geometrical statement into polynomial equations and inequations. At this stage a suitable kind of Geometry (Euclidean, Riemannian, etc...) is chosen. More precisely one has to point out the ground field k in which the polynomials take their values and express, in algebraic terms, the axioms for distances, areas, volumes, incidence etc...

The construction of the ground field is a very classical problem and was first studied by Hilbert ([17]). It is one of the principal motivations of the more recent studies on the so called algebra of the invariants. It is out of the aim of this paper to discuss this topic. A very modern and accessible introduction to it is [7], and a more classical one can be found in [20]. An account of this problem can also be found in [30].

A more practical aspect of this translation from Geometry to Algebra is the choice of the system of coordinates. It is in this stage that it is possible, for an algebraic prover to take advantage of the symmetries of the problem under consideration.

One of the possible formulations of the algebraic counter part of a geometric statement is the following:

Problem 1. Let k be a field and let $h_1(x_1, \ldots, x_n), \ldots, h_m(x_1, \ldots, x_n)$, $t(x_1, \ldots, x_n)$ polynomials with coefficients in k, said respectively *hypothesis polynomials* and *thesis polynomial.* It is requested to determine algorithmically if there exists a polynomial $d(x_1, \ldots, x_n)$ verifying the following properties:

i) d does not divide t;

ii) d is not in the ideal generated by the h_i's in the ring $k[x_1, \ldots, x_n]$;

iii) there exist polynomials $g_1, \ldots, g_m$ and an integer s such that

$$dt = \sum_{i=1}^{m} g_i h_i$$

Problem 1 can be restated in more geometrical terms as follows:

Problem 1'. Let H be an algebraic variety in k^n, called the hypothesis variety, containing all the points representing the configurations for which the hypothesis holds. Let T be the hypersurface of k^n described by the thesis polynomial containing all the configurations for which the thesis holds. It is requested to determine algorithmically if some open subset of H (in the sense of the Zariski topology on k^n) is contained in T.

These two equivalent formulations of the problem of automated theorem proving in Elementary Geometry are due to Wu ([32]). He calls it the problem of the mechanization of Elementary Geometry. They are the translation in rigorous algebraic terms of the concept of validity for a geometrical statement as currently accepted. It is important, in fact, to remark that most Geometry theorems are *valid* in the Logic sense only under the assumption of suitable non-degeneracy conditions.

For example the statement «there exists one and only one circle circumscribed to a given triangle» is valid only if the word «triangle» is used in the strict sense of a non-degenerated triangle, i.e., such that the vertices do not belong to the same line.

The non-degeneracy conditions are generally not easy to find as in the example reported above, and hence it is not realistic to require the user to provide them together with the statement to be proved. Hence together with the goal of producing a proof of a given theorem the prover has to determine a set of non-degeneracy conditions. They are algebraically represented by the polynomial d in Problem 1.

This setting of the problem is not, however, free from difficulties. It leaves, in fact, some arbitrariness in accepting a theorem as valid in «some» sense or in rejecting it as false or with not completely determined hypotheses. This problem is not merely theoretical, indeed it is possible to find several examples of statements which can be accepted by one of the provers described below and rejected as false or incomplete by another one. Moreover there is not a generally accepted criterion among the researchers in this area (see for example the remarks in [7] and the final remarks in [4]).

Another more technical difficulty is shown by the following example: (see [7])

EXAMPLE. *Let ABC be a triangle, and let BE be the perpendicular to the edge AC from B. Suppose the coordinates of these points are assigned in the following way:*

$$A \equiv (0,0);\ B \equiv (x_1,0),\ C \equiv (x_4,x_5);\ E \equiv (x_2,x_3).$$

The hypothesis BE normal to AC is expressed by the polynomial

$$h_1 = x_1x_4 - x_2x_4 - x_3x_5 = 0.$$

The hypothesis «E belongs to the segment AC» is expressed by the polynomial

$$h_2 = x_2x_5 - x_3x_4 = 0.$$

Finally the statement «ABC is isosceles» can be translated into

$$t = x_1^2 - 2x_1x_4 = 0.$$

It is easy to verify that Problem 1 in this case has a solution with $d = x_1 x_4 x_5$.

Hence if no attention is paid we obtain the paradoxical conclusion that the statement «Any triangle is isosceles» is a valid theorem!

This problem can be avoided is several ways. Chou, giving this example, points out that a possible solution is to examine the geometrical meaning of the polynomial d. But this cannot be, in general, done automatically, and it is sometimes a very hard task also for the user. Other researchers ([4], [5]) give the following specialization of Problem 1' as a possible solution:

Problem 2. Let H be an algebraic variety in k^n, called the hypothesis variety, containing all the points representing the configurations for which the hypothesis holds. Let T be the hypersurface of k^n described by the thesis polynomial containing all the configurations for which the thesis holds. It is requested to determine algorithmically if some irreducible component of H, of maximal dimension, is contained in T.

This formulation has the advantage of being completely mechanizable, and has been also successfully generalized (at least from a theoretical point of view) to Differential Geometry ([5]). On the other hand under this formulation some statement can be rejected as false even in the case in which some minor additional check to detect this anomalous situations is necessary.

A final problem that most of the algebraic provers have to face is that in the theory of polynomial rings over a real field it does not exist an equivalent of Hilbert Nullstellensatz which would simplify remarkably the algorithmic treatment. For this reason some algebraic provers work only over algebraically closed fields.

This condition can be sometimes misleading since a theorem can be true over the real field, but false in the complex field.

This causes serious consequences in methods based on the formulation given in Problem 2 above. Indeed it is possible that the irreducible components of maximal dimension mentioned in such formulation lie entirely in the complex part of the affine space.

Notice that the problem of the ground field is easily avoided by the «logical» prover, but these are generally not able to cope with the non-degeneracy conditions. As said above it is unrealistic to require a complete determination of such conditions by the user. A solution could be to use deduction rules less strict than the rules currently employed, but it is no very clear how these rules have to be formulated.

To solve these technical problems some of the current provers are strongly interactive. On the other hand this solution seems not to be theoretically acceptable because of its not completely algorithmic nature.

3. Wu's method.

In 1978 Wu WenTsun proposed an algorithm for automated theorem proving in Geometry based on the elimination procedure discovered by J.F. Ritt ([25]).

In what follows k will denote an algebraically closed field. A polynomial f in $k[x_1, \dots, x_n]$ is said to be of class i iff i is the maximum index such that f has a positive degree in x_i. The class of the elements of k is zero. If f is of class i the coefficient of the x_i of maximum degree (which is a polynomial in $k[x_1, \dots, x_{i-1}]$) is said to be the *initial* of the polynomial f and is denoted by $In(f)$. It is possible to order the polynomials of $k[x_1, \dots, x_n]$ in the following way: $f < g$ if g is of class greater than f or if f and g are of the same class i and $\deg_{x_i}(g) > \deg_{x_i}(f)$. f and g are otherwise not comparable.

If f and g are two polynomials of class respectively i and j, with $i < j$, or such that $i = j$ and the degree in x_i of f is less than the degree of g, then it is possible, using the Euclid algorithm over $k(x_1, \dots, x_{i-1})[x_i]$ to find polynomials q and r with $\deg_{x_i}(r) < \deg_{x_i}(f)$

such that

$$In(f)^{\alpha} g = qf + r$$

with α bounded by $\deg_{x_i}(f) - \deg_{x_i}(g) + 1$. The polynomial r is said the pseudo-remainder of g with respect to f, and it is denoted by $\mathrm{prem}(g, f)$. This operation is called *pseudo-division.*

A polynomial f is said to be reduced with respect to another polynomial g of class i if $\deg_{x_i}(f) < \deg_{x_i}(g)$. A set of polynomials $\{f_1, \ldots, f_r\}$ is said to be an *ascending set* (AS) if $r = 1$ and f_1 is in k, or if the following conditions are satisfied:

i) f_1 is not in k;

ii) $\mathrm{class}(f_1) < \mathrm{class}(f_2) < \ldots < \mathrm{class}(f_r)$;

iii) f_j is reduced with respect to f_i for all the pairs i, j with $1 \leq i < j \leq r$.

It is possible to define a generalization of the operation of pseudo division of a polynomial g with respect to an ascending set $\{f_1, \ldots, f_r\}$. By iterating the operation of pseudo division on g the following formula is produced:

$$In(f_1)^{\alpha_1} \ldots In(f_r)^{\alpha_r} g = q_1 f_1 + \ldots + q_r f_r + r.$$

and the polynomial r is called the *pseudo remainder* of g with respect to the ascending set $\{f_1, \ldots, f_r\}$.

It is possible to order the set of all the ascending sets. An ascending set A is said to be of rank less than another ascending set B if one of the following conditions holds:

1. Going up in the ascending order of indices there exists an index j such that the j-th polynomial in B is greater, according to the ordering described before, than the j-th polynomial in A.

2. All polynomials in B are not comparable with the corresponding elements of A but A has cardinality greater than B.

Plainly the relation $A \sim B$ iff A, B are not comparable is an equivalence relation. The set of equivalence classes is well-ordered by 1. and 2. ([25], [30]). Hence the collection of all ascending sets formed by polynomials of an ideal I has minimal elements. These minimal elements are said *characteristic sets* (CS), of the ideal;

An ascending set $\{f_1, \ldots, f_r\}$ is said to be irreducible if every polynomial f_i is irreducible, as polynomial in x_i, over the quotient field of the polynomial ring $k[x_1, \ldots, x_n]/(f_{i-1}, \ldots, f_1)$. This condition can be tested algorithmically (but no efficient algorithm is known). The importance of irreducible ascending sets is shown by the following proposition:

PROPOSITION 1. *Let A be a characteristic set for the ideal I. Then A is irreducible if and only if I is prime.*

The main properties of characteristic sets are summarized in the following proposition:

PROPOSITION 2. *Let $I = (f_1, \ldots, f_r)$ be an ideal in $k[x_1, \ldots, x_n]$ and let $\{h_1, \ldots, h_m\}$ be a characteristic set for I. Let J be the ideal generated by the h_i's and H^∞ the multiplicative set of all the products of the initials of the h_i's. The following properties hold:*

1. *$J \subseteq I \subseteq J : H^\infty$;*
2. *If $Z(I)$ denotes the set of the zeros of the ideal I then it results:*
$$Z(J) \setminus \bigcup_{i=1}^{m} Z(In(h_i)) \subseteq Z(I) \subseteq Z(J)$$
and the inclusions can be strict.
3. *If a polynomial f has pseudo remainder zero with respect to the characteristic set then it belongs to $J : H^\infty$. If the characteristic set is irreducible the converse is also true.*
4. *$dim(Z(J : H^\infty)) \geq n - m$.*
5. *If I is a prime ideal $H^\infty \cap I = \emptyset$, and then $I = J : H^\infty$.*

It is clear from this proposition that a *CS* carries enough information about its ideal only if the ideal is prime. Characteristic sets are not of easy computation. They can be computed either from a Gröbner basis of the ideal or in a direct way using the (optimal) algorithm in [11], or [12]. A more direct algorithm due to Ritt ([25]), and introduced in the authomated theorem proving by Wu ([30]) is based on successive computations of pseudo remainders. Unfortunately the set of polynomials produced by this algorithm generally is not a characteristic set of the assigned ideal, but it is still useful in automated theorem proving. In fact for these «Wu sets» proposition 2 above still holds.

An account of the Wu-Ritt algorithm can be found in [25], [30] and [11]. It is important to remark that in [30] the term «characteristic set» is applied to any ascending set with the properties listed in the above proposition 2, but classically this term is reserved only to minimal ascending sets.

To verify if a given polynomial g is zero over a variety one needs to have an algorithmic test to check if the ideal of the variety is prime, and if not to compute a suitable decomposition. In automated theorem proving one needs simply to verify if g is zero on an open subset of the variety, therefore the complete decomposition of the variety in its irreducible components is rarely necessary.

Wu's method to prove theorems in Geometry, can be now easily described. Starting from the polynomials representing the hypotheses one computes a characteristic set or a Wu set of this ideal. The second step is to compute the pseudo remainder of the thesis polynomial with respect to this set. If it is zero one can conclude that the theorem is true under the non-degeneracy conditions given by the initials of the characteristic set. If the pseudo remainder is not zero, no conclusion can be taken unless the irreducibility of the characteristic set is knwon. If this is the case the theorem can be rejected as false. Otherwise factorization is needed.

Wu's method has been implemented by Wu and his students in China and by Chou at Austin, Texas. This last implementation is very well documented in [7] and it includes also a part relative to the factorization of squared polynomials. The theorems verified by Chou's prover constitute the largest collections of mechanically proved theorems available. They represent the more explicit demonstration of the power and the elegance of Wu's method for automated theorem proving.

Finally Wu's method has been applied successfully to elementary differential geometry ([32]), by generalizing the Wu-Ritt procedure to differential algebra.

4. Methods based on Buchberger's algorithm and on the computation of the dimension.

After the success obtained by Wu's method several researchers tried to apply Buchberger's algorithm ([2], [3]) to the problem of mechanical theorem proving. This algorithm, which computes the so called Gröbner, or standard, bases for an ideal, is the main tool for the so called Computer Algebra.

Gröbner bases have ben applied to many different problems. For example it is possible to develop a complete decision procedure for the first order theory of algebraically closed fields entirely based on Buchberger's algorithm ([10]) and hence, at least theoretically, it is possible to prove theorems using Gröbner bases.

More specialized approaches have been, however, suggested by Kutzler and Stifter ([22], [23]), Kapur ([21]), Winkler ([28]) and by Carrà and Gallo ([4]). In this section it will be reported a brief account of the methods due to Kutzler and Stifter, Kapur, and Carrà and Gallo.

Kutzler and Stifter proposed the following version of the problem of automated theorem proving in geometry:

Problem 1". Let I be an ideal in the ring $k[x_1, \dots, x_n]$, and suppose that $I \cap k[x_1, \dots, x_d] = \emptyset$, i.e., the variables $x_1, \dots, x_d$ are independent with respect to I. Given a polynomial g (the conclusion) it is requested to decide algorithmically if there exists a polynomial s in $k[x_1, \ldots, x_d]$ such that gs is in the radical of the ideal I.

The solution proposed by Kutzler and Stifter, called RED-algorithm, is the following: compute, using Buchberger's algorithm, a Gröbner basis for the ideal $Ik(x_1, \ldots, x_d)[x_{d+1}, \ldots, x_n]$. If this extended ideal is equal to all the ring, it means that the first d variables are not really independent and another set of independent variables has to be tried, otherwise one simply reduces g with the rewrite rules given from the computed Gröbner basis. If the result of this reduction is zero the theorem is confirmed, otherwise is rejected.

This solution has several drawbacks. A guess has to be done for the set of independent variables, and this can require an exponential number of tentatives, but generally an analysis of the geometrical problem under consideration is helpful. Another difficulty is that one has to carry on a computation on the field $k(x_1, \ldots, x_d)$, and this increases the complexity of the method. Finally the RED-algorithm simply says if there exist a polynomial s, as in problem 1" such that gs is in I, but it does not give any information about the radical.

On the other hand the RED-algorithm does not make use of the Hilbert Nullstellensatz, and hence it can be successfully applied to algebraically non-closed fields.

In order to present an interesting variation of the RED-algorithm, called PRED-algorithm, Kutzler and Stifter introduce the concept of *u-pseudo reduction* with respect to some ordering of the monomials. $lc_k(q)$ denotes the coefficient of $lt(q)$, over the field k and $lm(q) = lc(q)lt(q)$. Let p, q, r be polynomials in $k[u_1, \ldots, u_d, x_{n-d+1}, \ldots, x_n]$. r u-pseudo reduces to p modulo q if the polynomial $lc_{k(u_1,\ldots,u_d)}$ $(q)r$ reduces to p modulo the rewriting rule $lm(q) \to q - lm(q)$. It is clear that the u-pseudo reduction is a kind of simulation, in term of rewrite rules, of the pseudo reduction in the Wu's sense, with the extra condition that the admissible «initials» here are polynomials only in the u_i's.

The PRED-algorithm can be now described as follows: compute a Gröbner basis for the ideal of the hypotheses, with respect to an ordering in which the variables u_i's are less than the others. As in RED the u_i's are guessed independent. If the computed basis contains a polinomial only in the u_i's another guess for the set of independent variables is necessary. If not the thesis polynomial, g, is u-pseudo reduced using the relations in the Gröbner basis. The theorem is confirmed if and only if g u-pseudo reduces to zero.

The advantage of this method is to avoid the computation on the field of rational functions, but it has the same other drawbacks as the RED-algorithm. It can be successfully applied also over algebraically non-closed fields. The approach of Winkler ([28]), which will not be reported here, tries to avoid the problems of the proposed methods by considering the problem of testing the radical membership for the thesis polynomial instead of the simple membership.

Kutzler and Stifter have conduct many experiments with their algorithms. They compared them also with Wu's algorithm. From the statistics they produce it is clear that their method is efficient and reliable as well as Wu's method.

Kapur's methods use Hilbert Nullstellensatz and the so called Rabinowitzch's trick to test the radical membership and, for this reason they are valid only for algebraically closed fields.

The first method proposed by Kapur assumes that the hypotheses of the geometrical statement are completely determined, i.e., the non-degeneracy conditions are given. Hence the hypotheses are expressed by polynomial equalities $h_1 = 0, \ldots, h_m = 0$, and by some polynomial inequalities $s_1 \neq 0, \ldots, s_t \neq 0$. This strong assumption leads to a very simple solution: one has just to test if the thesis polynomial g, is zero on the zeros of the h_i's which are not zeros of the s_j's. Using Rabinowitzch's trick this is done by testing if the ideal $J = (h_1, \ldots, h_m, z_1 s_1 - 1, \ldots, z_t s_t - 1, zg - 1)$ has zeros, where the z_i's are new auxiliary variables.

By Hilbert Nullstellensatz one has simply to test if 1 is in the ideal J. This can be done or by the algorithm of Hermann-Seidenberg ([16], [27]), using the improved bounds of Brownawell ([1]), or using Buchberger's algorithm.

Although the method is very simple and of immediate implementation (if an implementation of Buchberger's algorithm is available) it has the following drawbacks. As remarked in Section 2 the non-degeneracy conditions are generally very difficult to be determined «a priori». Moreover the introduction of many auxiliary variables increases the running time.

To prove theorems for which the non-degeneracy conditions are not known Kapur first proves the following simple result:

PROPOSITION 3. *Let $I = (h_1, \dots, h_m)$ be an ideal and g and p polynomials such that pg is in the radical of I. Then in any Gröbner basis of the ideal $J = (h_1, \dots, h_m, zg - 1)$, with respect to an ordering in which the auxiliary variable z is the smallest, there is a polynomial q with the same property of p.*

Using the above result Kapur suggests the following procedure: compute a Gröbner basis of the ideal J as in the proposition. If it is the unit ideal then the theorem holds without any non-degeneracy condition. Otherwise for every polynomial g_i in the Gröbner basis which does not belong to the ideal I of the hypotheses, test if $H_i = (h_1, \dots, h_m, zg_i - 1)$ contains 1. If this is the case then the theorem is false. Otherwise the theorem holds under the non-degeneracy condition $g_i \neq 0$.

An intensive comparison of the methods summarized above has been conducted by Kapur himself and by Chou ([7], [21]).

Finally the method developped by G. Carrà and Gallo ([4]) is based on the computation of the dimension of an algebraic variety (which can be done by a Gröbner basis computation). In contrast with the other methods of this section it can be generalized also to Differential Geometry even though no algorithm is known to construct a complete and confluent system of rewriting rules similar to Gröbner bases ([6]).

This method tries to avoid the difficulties presented by the other algebraic provers. Indeed they can accept as valid false propositions or reject statements which are valid under suitable non-degeneration hypotheses.

The only algebraic prover who takes care of this unsoundness is Wu's method which looks for the geometrical meaning of the non-degeneracy conditions produced by the prover (i.e., the initials in the set produced by the Wu-Ritt algorithm) to check if the proof produced is meaningful or not.

This solution is not completely satisfactory since it requires (computationally very expensive) factorization and an inverse translation from Algebra to Geometry which is not completely automated yet.

Carrà and Gallo distinguish the following validity of a geometrical statement to refine the formulation given in Section 1, Problem 2.

DEFINITION *Suppose that a geometrical statement is translated into a collection of polynomial equalities and inequalities for the hypotheses, and one polynomial equality for the thesis. The set of the points in the configuration space satisfying the hypotheses is called the hypothesis set, and the set of the points satisfying the thesis is called thesis set. These set are generally quasi-algebraic set, and have a decomposition in irreducible components according to the Zariski topology over the configuration space.*

A geometrical statement is said to be generically valid in a strong sense if all the components of the hypothesis set of maximal dimension are contained in the thesis set.

A geometrical statement is said to be *generically valid* if some component of the hypothesis set of maximal dimension is contained in the thesis set.

Strong validity, of course, implies validity but the converse is generally false. An example of a statement which is generically valid but not in a strong sense is the following:

EXAMPLE Let ABC be a triangle on the plane. Construct on the edges AC and BC two squares $ACDE$ and $BCFG$. Let M be the middle point of AB. Then DF is twice the lenght of CM.

The hypotheses define four three-dimensional components corresponding to the cases in which the squares are built inside or outside the triangle. The thesis is true only if both squares are built outside.

Therefore the statement is true only in the generic sense.

It is important to remark that it is generally very difficult to precise in an algebraic language relations as «inside» or «outside» which play a fundamental role here.

Chou has noticed, in [7] that it is possible to construct examples in which the dimension of the subvariety corresponding to degenerate case is greater than the dimension of the subvariety corresponding to the really interesting cases. This observation can be overcome by considering these statements as «incompletely described». Moreover the definitions of validity above are decidable in a totally mechanical way and do not require any mathematical expertise to the user.

Carrà-Gallo method uses the following property of quasi-algebraic sets:

PROPOSITION 4. *Let V be a quasi-algebraic set in $\mathbb{A}^n$ defined by the polynomial relations:*

$$f_1 = 0, \ldots, f_m = 0,\ g \neq 0.$$

Then V is isomorphic to the variety V' defined in $\mathbb{A}^{n+1}$ by

$$f_1 = 0, \ldots, f_m = 0, gT - 1 = 0$$

where T is a new auxiliary variable.

The proposition above allows to reduce the computation of the dimension of a quasi-algebraic set to the computation of the dimension of an algebraic set.

The algorithm to test strong validity of a geometrical statement can be described in the following way. First compute the dimension (i.e. the maximal dimension of the irreducible components) of the hypothesis set. Then compute the dimension of the set obtained by intersecting the hypothesis set and the complement of thesis set. If this second integer is less than the first one then all components of maximal dimension of the hypothesis set are contained in the thesis set, and the theorem is valid in the strong sense.

If the thesis is expressed by more than one polynomial then by multiplying them it is possible to reduce to the preceding case. However a separate test for each one of them gives information about which conclusion of the thesis has to be weakened to get a valid theorem.

The algorithm to test the generic validity is simpler than the method described above. It just requires to compute the dimension of the hypothesis set and the dimension of the set obtained from it by intersection with the thesis set.

If the two integers are equal then at least one maximal component of the hypothesis is contained in the thesis and the theorem is generically valid.

On the other hand if these two integers are not equal their difference gives an information on how many additional hypotheses one has to add to the statement in order to make it valid.

The core of the two methods described above is the computation of the dimension of an algebraic set. In 1987, when Carrà and Gallo first proposed their algorithms, two methos were known: the computation of the Hilbert polynomial of an ideal, and the analysis of a boolean matrix built from a Gröbner basis of an ideal. Both methods have the same worst case complexity than the computation of Gröbner bases, which is doubly exponential in the number of variables.

More recently it has been proved by several authors ([24], [9]) that the dimension can be calculated by simply exponential algorithms based on suitable versions of Hilbert Nullstellensatz (see [1], [11]). No study it is known about the average complexity of these methods. An experimentation of the Carrà-Gallo method has been done at the University of Catania and a report about it is in preparation.

Since algorithmic methods to compute the dimension of algebraic differential ideals are known ([6]) it has been possible to generalize these methods to the case of elementary differential geometry ([5]). The complexity of the available algorithms is however too high, and this generalization has just theoretical interest.

As for the algebraic case, one has to give a precise definition of when a differential geometry statement is considered valid. It is possible to proceed as above, but in the differential case one has to take care of the possibility that an irreducible differential algebraic set can contain another differential algebraic set, which is irreducible and of the same dimension. This justifies the following definition.

DEFINITION

i) Thesis follows in a generic strong sense of type 1 (gs1) from hypothesis if the differential algebraic set defined by the thesis contains all the irreducible components of maximal dimension of the hypothesis set.

ii) Thesis follows in a generic sense of type 1 (g1) from hypothesis if the differential algebraic set defined by the thesis contains some irreducible component of maximal dimension of the hypothesis set.

iii) Thesis follows in a generic strong sense of type 2 (gs2) from hypothesis if the differential algebraic set defined by the thesis contains, for all the irreducible components of maximal dimension of the hypothesis set, an irreducible subset of the same dimension.

iv) Thesis follows in a generic sense of type 2, (g2) from hypothesis if the differential algebraic set defined by the thesis contains, for some of the irreducible components of maximal dimension of the hypothesis set, an irreducible subset of the same dimension.

Using the same techniques described in the algebraic case (i.e. the computation of the dimensions of the hypothesis set, of the intersection of the hypothesis set with the complement of the thesis, and the intersection of the hypothesis set with the thesis set) it is possible to decide if a given statement is not valid according to any of the above definitions. If this is not the case then it is possible to check if the statement is valid according to definition i).

Nothing is known about the computational complexity of this method.

5. Proving by example.

One of the most peculiar characteristic of the Gelernter's prover, as reported in section 1, is the use of diagrams, i.e, ultimately, of examples, to decide some property that was out of the logical power of the prover.

Jiawei Hong ([18]) has proposed a prover which uses the same approach but in a more rigorous mathematically justified sense. Hong's «proving by example» method was originally designed to prove theorems for which hypotheses and thesis are expressible by polynomial equations, with degree at most two in each variable. Later Hong discovered a method to prove inequalities over the real numbers for a large class of functions, including polynomials ([18]). In this paper we will give an account only of the first algorithm.

Hong formalizes the class of geometrical statements that can be proved by his prover as follows. Any geometrical statement is composed by three basic elements:

- the choice of a finite number, say s, of arbitrary points;
- a sequence of constructions carried out starting from the initial points;
- an assertion about the equality of some of the points constructed in the previous steps.

Algebraically this leads to the introduction of $2s$ independent variables and of r variables which are dependent over the others, according to some polynomial relations. Hong assumes that the new points are constructed only by intersecting lines connecting two points and circles centered at one point and containing another point. This implies that the generated polynomial relations have degree at most two in each variable. Another consequence of these premises is that the polynomial relations have the form of a triangular system of the following kind:

$$f_1(u_1, \ldots, u_{2s}, x_1) = 0.$$

$$\ldots$$

$$f_{r-1}(u_1, \ldots, u_{2s}, x_1, \ldots, x_{r-1}) = 0$$

$$f_r(u_1, \ldots, u_{2s}, x_1, \ldots, x_{r-1}, x_r) = 0$$

To prove a theorem means to test if one or more polynomials are zero on the *(generic)* zeros of the variety defined by the equations above. Hong precises the adjective generic as follows.

Starting from the equations above one can (at least in theory) solve them in sequence, in such a way that the values taken by the x_i's are functions of the values taken by the u_i's (the parameters). In particular considering the bounds assumed on the degrees one is able to write down an explicit formula for such functions.

For each choice of the parameters, it is possible to build a tree in the following way. The root is a vertex labelled by the $2s$-tuple of the values chosen for the u_i's. It has one, two, or no children, if f_1 has one, two, zero or infinite solutions respectively as equation in x_1; If a solution can be computed (assuming that one is able to perform exact real arithmetic) it will label the newly created node. The same procedure applies recursively to the successive levels of the tree. If a node has no children it is labelled by an X (degenerate case). If the tree has nodes at the r-th level (i.e., it is possible to complete the construction) then for each of these node one has a complete set of values for the u_i's and the x_j's and then it is posisble to test the thesis for this $2s + r$-tuple of real numbers. An appropriate label is assigned to these terminal nodes if the thesis is true for them or not.

It is clear that the analysis of the tree gives all the informations one needs about the theorem in consideration (non-degeneracy conditions, validity, special cases, etc...). It is remarkable that the strucutre of the tree is independent of the decomposition in irreducible components of the hypothesis variety.

The trees constructed above are isomorphic for almost all choices of the parameters, and Hong proves the following result:

PROPOSITION 5. *There exists a polynomial $t(u_1, \dots, u_{2s})$ such that for any $2s$-tuple $(v_1, \dots, v_{2s})$ one of the following two facts is true:*

i) $t(v_1, \dots, v_{2s}) = 0$;

ii) the tree constructed starting from the values $(v_1, \dots, v_{2s})$ is isomorphic to the tree constructed by any other $2s$-tuple which is not a solution of the polynomial t.

The degree of the polynomial t can be computed from the degrees of polynomials in the hypothesis.

According to the above theorem it is possible to say that a choice of parameters is *generic* when it is not solution of the polynomial t. Thus if one is able to pick up an example satisfying such condition then the theorem can be proved by just verifying that example.

This idea gives an effective algorithm if one specifies how to realize a generic choice and how to cope with the problem raised by real arithmetic.

To solve the first problem Hong introduces the concept of *generic basis for a variety*. Let $P(i, p, t, v)$ be the set of all the polynomials in $k[u_1, \dots, u_{2s}, x_1, \dots, x_i]$ with total degree less than p, sum of the absolute values of the coefficients less than t, and degree in each of the variables less than v. The $2s$-tuple $(u_1, \dots, u_{2s})$ is said a basis for the point $(u_1, \dots, u_{2s}, x_1, \dots, x_i)$.

The $2s$-tuple $(u_1, \dots, u_{2s})$ is said a *generic basis* for the variety

A, with respect to $P(i,p,t,v)$ if the following property holds: for any polynomial f in $P(i,p,t,v)$ and any irreducible component of A, f is zero over this component if and only if it has a zero whose basis is $(u_1,\dots,u_{2s})$.

As a consequence of the above theorem Hong is able to give an explicit construction of the generic basis of the hypothesis variety. Suppose that the construction starts with s arbitrary points, and takes q steps. Let $p = c^q$ and $t = p^{2^p}$ with c an absolute constant. Then

$$z_1 = pt;$$

$$z_2 = (pt)^{1+p};$$

$$z_3 = (pt)^{1+p+p^2};$$

$$\dots$$

$$z_{2s} = (pt)^{1+p+p^2+\dots+p^{2s-1}};$$

is a generic basis for the variety determined by the hypotheses.

One has just to analyze the example built starting from this parameters to prove the theorem.

The last difficulty to be considered is how to work with real numbers, which require infinite precision, using just an approximate arithmetic? To answer this question Hong proves the following «gap theorem»:

PROPOSITION 6. *Let $\{a_n\}$ be a sequence of real numbers each one obtained from the preceding ones and from a finite number of parameters bounded by a constant M, by means of the operations of sum, product, division and square root computation. Then either a_i is zero or its absolute value is greater than M^{-c^i}, where c is a constant depending only on the number of the parameters.*

As a consequence of the theorem the precision one needs to test a geometrical statement as above is of d^{ls} digits, where d is some constant. Using this bound Hong is able to prove that his method has a parallel complexity of $O(\log^h(k^{ls}))$, where h an k are constants.

Hong's method is very original and elegant, but it presents some difficulties. The size of the constant involved in the algorithm seems to be very large (since it must keep into account the extreme instability of the roots of a polynomial as a function of the coefficients), moreover no complete implementation of the method has been reported, even though Hong claims that some new theorems have been proved by this algorithm. Finally the method seems not immediately extensible to the general case where the hypothesis can be expressed by arbitrary polynomials.

6. The probabilistic approach.

From what has been so far described the task of automatic theorem proving in Geometry appear to be computationally very expensive. It is therefore natural to ask for probabilistic algorithms. This approach will not give in general complete answers about the validity of a theorem, but it will at least classify it as a «certified» conjecture with a given high probability.

Altough probabilistic algorithms have been recently widely developed for several areas of Computer Science, not very much is known about the use of probabilistic algorithms in geometry theorem proving. The only method described in the literature is given by J.T. Schwartz [26].

His approach is based on an observation of P.J. Davis [8] who notes that geometric theorems can always be reduced to a single algebraic identity. This is for example an immediate consequence of both Wu's and Hong's methods.

Two main tasks arise:

a) Finding efficiently the algebraic identity $f \equiv 0$ which is equivalent to the theorem to be proved.

b) Verifying that algebraic identity $f \equiv 0$.

There are many methods to perform a) but there is no study on the efficiency of various strategies.

So we will concentrate on task b). This can either be performed by explicitely calculating the polynomial f, simplifying it and showing that it is zero. This is done in Wu's method. On the other hand in Hong's method f is not explicitely calculated but an extimation of the maximum size of the roots of the polynomial f is found. In this approach one tries to prove that $f \equiv 0$ by showing that f is zero on a point which exceeds the maximum allowed size.

A third method, [26] [33] consists of proving that f has too many roots by evaluating it on sufficiently many points using Hong's Gap theorem.

Schwartz suggests a probabilistic approach to geometry theorem proving by giving probabilistic algorithms for checking polynomial identities.

First he considers the case of integer arithmetic and notices that the following fact holds [26].

PROPOSITION 7. *Let $f(x_1, \ldots, x_n)$ with coefficients in a field k be a polynomial which is not identically zero. Let I be a subset of k such that the number $|I|$ of elements of I is greater than 2deg(f). Then in the set $I^n = I \times I \times \ldots \times I$ the polynomial f has at most $|I|^n/2$ roots.*

This suggests the following probabilistic test to check $f \equiv 0$:

(i) Chose I such that $|I| \geq$ 2deg(f)

(ii) Let m be an integer such that 2^{-m} is small enough (*for example* $m = 100$)

(iii) Select m elements in I^n at random.

(iv) If any of these elements is not a root of f then $f \not\equiv 0$. Otherwise if they are all zeros of f then $f \equiv 0$ with probability $1 - 2^{-m}$.

If the polynomial f is given in an implicit form then finding its explicit expression could be very expensive. In this case the probabilistic approach could be useful.

For example Schawrtz notes that if f is the Vandermonde determinant over say 100 variables the total degree is about 5000 and (hopeless) simplification by expansion will generate about 2^{5000} terms!

Using the above probabilistic test in order to guarantee an accuracy of 10^{-100} one needs to run about 60 random tests on an ipercube of $\mathbb{N}^{100}$ of side 250.000. This will involve no more than 10^8 arithmetic operations which can be performed in a few minutes.

In order to keep the size of the coefficients low one can use modular arithmetic.

Another, more challenging, approach is to use real arithmetic to verify $f \equiv 0$. In the case of polynomials in one variable Schwartz is able to give a probabilistic algorithm based on the following fact: (see [26]).

PROPOSITION 8. (Kayeka-Okada-Fekete-Szegö) *Let $f(x)$ be a polynomial with integer coefficients which is not identically zero on the reals. Then the measure of the set $\{x \in \mathbb{R} \mid |f(x)| < 1\}$ is not greater than 4.*

This result suggests the following real probabilistic test:

(i) Choose an interval I of measure greater than 8.

(ii) Let m be an integer such that 2^{-m} is small enough.

(iii) Select m elements randomly in I.

(iv) If any of these elements makes $|f|$ greater than zero then $f \not\equiv 0$.
Otherwise if they are all zeros (within the precision of the calculation) then $f \equiv 0$ with probability $1 - 2^{-m}$.

Unfortunately Schwartz is not able to give an analogous test for multivariate polynomials.

Finally it is worth noticing that sometimes it is not immediate to reduce a geometrical theorem to a single identity.

One can then try to reduce the theorem to an implication $f_1 = 0$ and $f_2 = 0 \Rightarrow f = 0$ with no more than two hypotheses and degree at most 4. In this case one can use the above probabilistic tests in connection with the classical elimination theory of Kronecker, to test if the various resultants are identically zero (see [30]).

In conclusion we can say that this probabilistic approach seems to be very promising but must be subject to more deep investigation.

REFERENCES

[1] Brownawell W.D., *Bounds for the degree in the Nullstellensatz,* Annals of Math **126**, (1987), 577-591.

[2] Buchberger B., *An algorithm for finding a basis for a residue class ring of a zero dimensional polynomial ideal,* PhD theis (1965) Innsbruck.

[3] Buchberger B., *Gröbner bases an algorithmic method in polynomial ideal theory,* in Recent trends in multidimensional system theory, ed. Bose (1985), Reidel, Ma.

[4] Carrà G., Gallo G., *A procedure to prove geometrical statements,* LNCS **356**, (1987) 141-150.

[5] Carrà G., Gallo G., *A procedure to prove statements in Differential Geometry,* To appear on J. of. Aut. Reasoning.

[6] Carrà G., *Gröbner bases and Differential Algebra,* LNCS **357** (1988).

[7] Chou S.C., *Mechanical Geometry Theorem Proving,* Reidel, Ma (1988).

[8] Davis P.J., *Proof, completeness, trascendentals, and sampling,* Journal of ACM **24** (1977), 298-310.

[9] Dickenstein A., Fitchas N., Giusti M., Sessa C., *The membership problem for unmixed polynomial ideals is solvable in sub exponential time,* preprint 1989.

[10] Ferro A., Gallo G., *Gröbner bases, Ritt's algorithm and decision procedures for algebraic theories,* LNCS **356**, (1987) 230-237.

[11] Gallo G., *La dimostrazione automatica in Geoemtria e questioni di complessità correlate,* PhD thesis Catania.

[12] Gallo G., Mishra B., *Some bounds on the degrees of a characteristic set,*

preprint.

[13] Gelernter H., *Realization of a Geometry theorem proving machine,* in Computers and Thought eds Feigenbaum and Feldmann 134-153 Mc GrowHill NY (1963).

[14] Gelernter H., Hansen J.R., Loveland D.W., *Empirical exploration of the Geometry theorem proving machine,* inComputers and Thought eds Feigenbaum and Feldmann McGrowHill NY (1963), 153-163.

[15] Gilmore P.C., *An examination of the Geometry theorem proving machine,* J. of A.I. **1**, (1970) 171-187.

[16] Hermann G., *Die Frage der endlick vielen Schritte in der Theorie der Polynomideale,* Math. Annales **95** (1926) 736-788.

[17] Hilbert D., *Grundenlagen der Geometrie,* 1889.

[18] Hong J., *Proving by example and Gap theorems,* Proc of the 27th Annual Symp. FOCS (1986) 107-116 Toronto. To appear on Journal of Autom. Reasoning.

[19] Hong J., *Proving inequalities by example,* Tech. Rep. Univ. of Chicago, Chicago Ill. (1988). To appear on Journal of Autom. Reasoning.

[20] Hodge D., Pedoe W., *Methods of Algebraic Geometry,* Academic Press 1956.

[21] Kapur D., *Geometry Theorem Proving using Gröbner bases,* J. of Symb. Comp. **2** (1986) 399-412.

[22] Kutzler B., Stifter S., *On the application of Buchberger's algorithm to automated geometry theorem proving,* J. of Symb. Comp. **2**, (1986), 389-398.

[23] Kutzler B., Stifter S., *Collection of computerized proofs of geometry theorems,* Tech. Rep. Univ.Linz 1986.

[24] Logar A., *A computational proof of the Noether's normalization lemma,* LNCS **357** (1988).

[25] Ritt J.F., *Differential Algebra,* AMS 1950.

[26] Schwartz J.T., *Fast Probabilistic Algorithms for verification of Polynomial Identities,* Journal of ACM, **27**, (1980) 701-717.

[27] Seidenberg A., *Constructions in Algebra,* Trans. AMS **197**, (1974) 273-313.

[28] Winkler F., *A geometrical decision algorithm based on the Grobner bases algorithm,* submitted to ISSAC-88 (1988).

[29] Wos L., Overbeek R., Lusk E., Boyle J., *Automated Reasoning,* Prentice Hall 1984.

[30] Wu W.T., *Toward mechanization of Geometry - Some comments on Hilbert's Grundenlagen der Geometrie,* Acta Math. Scientia **2**, (1982) 125-138.

[31] Wu W.T., *On the decision problem and the mechanization of theorem proving in elementary geometry,* in Automated Theorem Proving after 25 years, Bledsoe and Loveland EDS., Contemporary Math. **29** (1984) 235-242.

[32] Wu W.T., *Mechanical derivation of Newton's gravitational law,* from Kepler's laws To appear on Journal of Autom. Reasoning.

[33] Zhang J., Yang L., Deng M., *The Parallel Numerical Method of Mechanical Theorem Proving,* Report IM530 · Academia Sinica. (Oct. 1988).

Alfredo Ferro
Università di Catania
Dipartimento di Matematica
Catania

Giovanni Gallo
Department of Computer Science
Courant Institute of Mathematical Sciences
New York University, USA