The Idea of a Diagram

Desmond Fearnley-Sander*
Department of Mathematics
University of Tasmania

. . . although [mathematicians] make use of the visible forms and reason about them, they are thinking not of these, but of the ideals which they resemble; not of the figures which they draw, but of the absolute square and the absolute diameter, and so on—the forms which they draw or make are converted by them into images, but they are really seeking to behold the things in themselves, which can only be seen with the eye of the mind.

Plato, *The Republic*, Book VI.

A reasonable programming language for geometry would

- implement some kind of uniform representation of geometric entities;
- transparently organize geometric entities into natural hierarchical structures;
- be capable of deducing properties of geometric entities from their specification alone;
- be able to compute geometric quantities such as volume;
- embody a formalism which is terse, natural and preferably resembles traditional mathematical usage;
- transparently avoid coordinates whenever possible, but use them when necessary;
- make use of the natural geometric typing that, for example, distinguishes points from vectors;
- handle transparently the symmetries inherent in typical geometric propositions;

*This paper appeared in *Resolution of Equations in Algebraic Structures Volume 1*, ed. Hassan Ait-Kaci and Maurice Nivat, Academic Press, 1989, 127-150.

- be powered by a strong theorem prover, which in the course of attempting a proof can uncover more useful information than merely whether the attempt succeeds or fails;
- cope painlessly with the assumptions of genericity that are often implicit in geometric statements;
- be suited to an interactive implementation (in accordance with the natural way of applying geometry, for example, in CAAD);
- interface smoothly with graphic and logic languages.

The critical problem in the development of such a language is the first of these items—how to represent geometric knowledge, and, in particular, capture the mathematician's notion of a diagram. This paper focusses on a solution of this problem. It is presented in terms of a particular formalism, but will be seen to be meaningful for other algebraic formalisms. Examples are given of how theorems of affine geometry—properties of diagrams—may be proved by equational reasoning. The ideas have been partially implemented (see [12]).

1 Grassmann algebras

The basic theory of Grassmann algebras, on which the formalism is based, is given (with minor differences in terminology from the present paper) in [10].

A *Grassmann algebra* is a ring which is generated by (the union of) disjoint subsets $\mathcal{K}$ and $\mathcal{P}$ such that

GA1 $\mathcal{K}$ is a field (under the ring operations);

GA2 $\mathcal{P}$ is an affine space over $\mathcal{K}$ (under the ring operations);

GA3 $aA = Aa$ for every $a \in \mathcal{K}$, $A \in \mathcal{P}$;

GA4 $BA = -AB$ for every $A, B \in \mathcal{P}$.

Here, **GA2** means that

$$A, B \in \mathcal{P}, a, b \in \mathcal{K} \text{ and } a + b = 1 \Rightarrow aA + bB \in \mathcal{P}.$$

An important consequence of **GA4**, provided the characteristic of $\mathcal{K}$ is not 2, is

GA5 $A^2 = 0$ for every $A \in \mathcal{P}$.

It is easy to show that elements of $\mathcal{V} = \mathcal{P} - \mathcal{P}$ also have the properties **GA4** and **GA5**.

If there is a maximal finite set of elements of $\mathcal{P}$ which is linearly independent over $\mathcal{K}$, then the Grassmann algebra is said to be *finite-dimensional*, and the

cardinality, less 1, of such a set is called its *dimension*. For a given field $\mathcal{K}$ and natural number n, there is, up to isomorphism, precisely one Grassmann algebra of dimension n.

In the present paper we take $\mathcal{K}$ to be the field of rational numbers.

2 Grassmann geometry

Grassmann geometry is the language in which one speaks about Grassmann algebras. The *expressions* of Grassmann geometry are the terms generated by

point variables $A, B, C, \ldots$

vector variables $U, V, W, \ldots$

point constant O

vector constants $X = X_1, Y = X_2, Z = X_3, X_4, \ldots$

scalar constants $0, 1, -1, \frac{1}{2}, \ldots$ (rational numbers)

operation symbols $+, *, -$ (of arities 2,2,1)

We use infix notation for addition and multiplication, and usually indicate multiplication by juxtaposition. An *atom* is a variable or constant. A *ground* expression is one all of whose atoms are constants.

Rewrite rules corresponding to the Grassmann algebra axioms (together with "built-in" addition and multiplication tables for scalar constants) may be used to transform any expression into a sum of products of atoms. Indeed by ordering the generating atoms a unique *normal form* is obtainable. The set of normal expressions may be endowed with an algebraic structure having the Grassmann algebra signature in the natural way (by defining, for example, the sum of two normal expressions exp_1 and exp_2 to be the normal form of the expression $\mathsf{exp}_1 + \mathsf{exp}_2$). We call this the UNIVERSE. We refer to an expression in normal form as a *polynomial*, and to the equivalence class of all expressions having a common normal form as a *quantity*. A polynomial is a sum of products of atoms, and from now on we call these products the *terms* of the polynomial (and refer to elements of the term algebra as expressions).

The *degree* of a term is the number of point and vector atoms occuring in it. If all the terms in the normal form of an expression have the same degree, then the expression is said to be *homogeneous*.

An expression of degree zero is called a *scalar*. A first degree homogeneous expression is called a *point*, if the sum of the coefficients of all the point atoms in its normal form is 1, and a *vector*, if this sum is 0. We denote by NUMBERS, POINTS and VECTORS the sets, respectively, of all scalar constants, of all points and of all vectors. This typing may be extended in a natural way to higher degree terms.

Taking $\mathcal{K}$ to be NUMBERS, $\mathcal{P}$ to be POINTS and $\mathcal{V}$ to be VECTORS, the UNIVERSE is a Grassmann algebra. For each natural number n, the subalgebra WORLD_n generated by $O, X_1, X_2, \ldots, X_n$ is a Grassmann algebra of dimension n; we call it the *world of dimension* n. By convention, WORLD_0 is NUMBERS.

Although it is the UNIVERSE that we are concerned with, in practice it is handy to allow its elements to be represented in non-normal form since this makes interpretation easier—for example $(B-A)(D-C)$ is immediately recognizable as a product of two vectors, while its normal form $BD-BC-AD+AC$ is not.

A *configuration of type* (k, m) is a finite sequence of k points followed by m vectors—that is, an element of $\mathsf{POINTS}^k \times \mathsf{VECTORS}^m$. A configuration is called *atomic* (or *variable* or *constant*) if all its elements are atoms (or, respectively, variables or constants). This notion of configuration, which suffices for the current purpose, can usefully be generalized.

3 Equations

An *equation* is simply a monic homogeneous polynomial *poly*. It is convenient to allow it to be represented by any pair of expressions $(\mathsf{exp}_1, \mathsf{exp}_2)$ such that $\mathsf{exp}_1 - \mathsf{exp}_2$ represents the same quantity as some non-zero constant scalar multiple of poly; and to write such a pair in the familiar form

$$\mathsf{exp}_1 = \mathsf{exp}_2.$$

Particularly useful representations are the *polynomial representation*

$$\mathsf{poly} = 0$$

and the *head-body representation*

$$\mathsf{head} = \mathsf{body},$$

where head is the maximal product term of *poly*.

A *labelled object* is a homomorphism of the UNIVERSE into itself which maps finitely many atomic variables (say, var_1, var_2,...,var_k) to constants (val_1, val_2,...,val_k) and leaves fixed all other variables ; it is completely determined by the set of equations

$$\mathsf{var}_1 = \mathsf{val}_1, \mathsf{var}_2 = \mathsf{val}_2, \ldots, \mathsf{var}_k = \mathsf{val}_k.$$

We denote by LABELLINGS the set of all labelled objects.

If σ is a labelled object, we call $\sigma(\mathsf{exp})$ the *value of the expression exp on the labelled object*. If $\sigma(\mathsf{exp}) = 0$, we say that the *expression vanishes on the labelled object*.

Example

The equations

$$A = O$$

$$B = O + X$$

$$C = O + X + Y$$

$$D = O + Y$$

describe a labelled object. The expression $ABC - ACD$ vanishes on this labelled object. The expressions $(B - A)(C - A)$ and $AB + BC + CA$ have the same normal form; and on this labelled object, they have the same value, namely XY. □

In general, expressions exp_1 and exp_2 have the same normal form (that is, represent the same quantity) if and only if they have the same value on every labelled object. Thus quantities have well-defined values on labelled objects.

An object labelling σ is said to *satisfy* (or be a *solution* of) an equation

$$\mathsf{exp}_1 = \mathsf{exp}_2$$

if exp_1 and exp_2 have the same value on σ .

Example

The labelled object of the previous example satisfies the equations

$$ABC = ACD$$

and

$$C - A = X + Y. \ \square$$

We shall often need to speak of sets of equations, and coin the term *description* as an abbreviation. We denote by DESCRIPTIONS the set of all descriptions.

A labelled object is said to *satisfy* a description if it satisfies every equation in it. For a description eqns, we denote by SOLUTIONS(eqns) the set of all labelled objects that satisfy it.

A description eqns is called

- *true* or *valid* if SOLUTIONS(eqns) = LABELLINGS (that is, if every labelled object satisfies it), and invalid otherwise;
- *unsatisfiable* or *false* if SOLUTIONS(eqns) = $\emptyset$ (that is, if there exists no labelled object that satisfies it), and *satisfiable* otherwise.

If eqns consists of ground equations, then eqns is necessarily either true or false.

Type-checking alone can establish that individual equations are unsatisfiable; hence an implementation can reject attempts to input equations like $C-B+A=0$ (because 0 is a vector while $C-B+A$ is not) and $AB=XY$ (because AB and XY are of different types).

4 Theorems

An *elementary assertion* is a pair of descriptions (hyps, conc); it is called an *elementary theorem*, and we say that hyps *semantically implies* conc, and write

$$\mathsf{hyps} \models \mathsf{conc},$$

if SOLUTIONS(hyps) $\subseteq$ SOLUTIONS(conc) (that is, if every labelled object that satisfies hyps satisfies conc). Clearly hyps semantically implies conc if and only if hyps semantically implies each equation of conc. We say that an equation eqn is a *consequence* of a description hyps if hyps $\models$ {eqn}. We denote by CON(hyps) the set of all consequences of hyps. Then CON is a closure operation on the space of all descriptions.

Example

It will be helpful to represent a typical assertion like $(M=\frac{1}{2}(A+B), AM=MB)$ in the more expansive notation:

hyp $\quad M=\frac{1}{2}(A+B)$
conc $\quad AM=MB$

This is actually a theorem, as we show below. □

Two descriptions eqns_1 and eqns_2 are called *semantically equivalent* if

$$\mathsf{eqns}_1 \models \mathsf{eqns}_2 \text{ and } \mathsf{eqns}_2 \models \mathsf{eqns}_1$$

(that is, if the sets of labelled objects that satisfy them are identical). Note that $\{A-B+C-D=0\}$ and $\{P-Q+R-S=0\}$ are *not* semantically equivalent.

For a set of labelled objects objs, the set of quantities

$$\begin{aligned}\mathsf{IDEAL(objs)} &= \{\mathsf{poly} : \sigma \in \mathsf{objs} \Rightarrow \sigma(\mathsf{poly})=0\}\\ &= \{\mathsf{poly} : \text{ every } \sigma \in \mathsf{objs} \text{ satisfies } \mathsf{poly}=0\}\end{aligned}$$

is an ideal in the UNIVERSE. In this ring-theoretic notation,

$$\mathsf{CON(eqns)} = \mathsf{IDEAL(SOLUTIONS(eqns))},$$

and the ideal generated by eqns is a subset (usually proper) of this ideal.

5 Diagrams

The fundamental entities of geometry are *diagrams*. We now define this simple but abstract notion.

A *diagram definition* is a pair $(\mathsf{varconfig}, \mathsf{eqns})$, consisting of a variable configuration $\mathsf{varconfig} = (\mathsf{pt}_1, \ldots, \mathsf{pt}_k, \mathsf{vec}_1, \ldots, \mathsf{vec}_m)$, say, and a description eqns. The *diagram* it defines is the map

$$\mathsf{POINTS}^k \times \mathsf{VECTORS}^m \to \mathsf{DESCRIPTIONS},$$

whose value at any configuration

$$(\mathsf{ptexp}_1, \ldots, \mathsf{ptexp}_k, \mathsf{vecexp}_1, \ldots, \mathsf{vecexp}_m) \in \mathsf{POINTS}^k \times \mathsf{VECTORS}^m$$

is the set of all consequences of the description obtained by replacing each occurence of pt_1 in eqns by ptexp_1, each occurrence of pt_2 by ptexp_2, and so on.

We use an obvious notation to denote the assignment of a name to a diagram. For example,

$$\mathsf{def} \quad \underline{\mathsf{parallel}}(V, W) \longmapsto VW = 0,$$

means that $\underline{\mathsf{parallel}}$ is the name given to the diagram defined by the pair $((V, W), \{VW = 0\})$.

A *call* of the diagram $\underline{\mathsf{dgm}}$ is a pair consisting of the name $\underline{\mathsf{dgm}}$ and a configuration of the appropriate type; and, for example, the notation

$$\underline{\mathsf{parallel}}(U + V, W) \longmapsto (U + V)W = 0$$

(lacking the annotation def) means that the value of the diagram $\underline{\mathsf{parallel}}$ (previously defined) at the configuration $(U + V, W)$ is the set $\mathsf{CON}\{(U + V)W = 0\}$. In a diagram call, the type of the configuration to which the diagram is applied must match the type that was (implicitly) declared in the definition.

We may also define a diagram in terms of other diagrams, as in the example

$$\mathsf{def} \quad \underline{\mathsf{vertical}}(A, B) \longmapsto \underline{\mathsf{parallel}}(B - A, Y).$$

(The symbol $\longmapsto$ may be read as "rewrites to".) This allows economical storage of diagram definitions, and imposes a natural hierarchical structure on the set of diagrams. For example:

$$\mathsf{def} \quad \underline{\mathsf{parallelogram}}(A, B, C, D) \longmapsto A - B + C - D = 0$$

$$\mathsf{def} \quad \underline{\mathsf{wall}}(A, B, C, D) \longmapsto \underline{\mathsf{parallelogram}}(A, B, C, D), \; \underline{\mathsf{horizontal}}(A, B), \; \underline{\mathsf{vertical}}(B, C)$$

$$\mathsf{def} \quad \underline{\mathsf{door}}(A, B, C, D) \longmapsto \underline{\mathsf{parallelogram}}(A, B, C, D),$$

$$B = A + 3X,$$
$$C = B + 7Y$$

def $\quad \underline{\mathsf{house}}(A, B, C, D, E, F, G, H, K) \longmapsto \underline{\mathsf{wall}}(A, B, C, D),$
$\underline{\mathsf{vertical}}(\frac{1}{2}(A + B), E),$
$E - \frac{1}{2}(C + D) = \frac{1}{2}(D - A),$
$\underline{\mathsf{door}}(F, G, H, K),$
$\frac{1}{2}(F + G) = \frac{1}{2}(A + B)$

While a diagram definition is a finite specification of a function, we may also conceive it in a procedural way—to show, for example, that a configuration (A, B) is vertical, show that the configuration $(B - A, Y)$ is parallel.

6 Realizations

We say that a configuration config *realizes* a diagram dgm if dgm(config) is true. If config is constant then dgm(config) is necessarily either true or false, and so we may even view a diagram as a predicate.

A diagram, such as the one defined by

def $\quad \underline{\mathsf{fred's\_front\_door}}(A, B, C, D) \longmapsto A = O,$
$B = O + 3X,$
$C = O + 3X + 7Y,$
$D = O + 7Y,$

whose description is a labelled object, is called an *object*. There is an obvious one-to-one correspondence between objects and constant configurations, and it is sometimes convenient to slur the distinction between the two concepts. An object may be viewed as belonging to any world that contains all its constants. For example fred's_front_door belongs to WORLD_2 (and to WORLD_3).

Example

While adding constraints to existing diagrams is a natural mode of definition it is not the only way to proceed. For example, new_door defined by

def $\quad \underline{\mathsf{new\_door}}(A, B, C, D, V)$
$\longmapsto \underline{\mathsf{fred's\_front\_door}}(A - V, B - V, C - V, D - V)$
$\longmapsto A = O + V,$
$\longmapsto B = O + 3X + V,$
$\longmapsto C = O + 3X + 7Y + V,$
$\longmapsto D = O + 7Y + V,$

is not an object. The realizations of new_door are precisely the translates of fred's_front_door. □

7 Properties

We say a diagram call $(\underline{\mathsf{dgm}_1}, \mathsf{config}_1)$ *has the property* $(\underline{\mathsf{dgm}_2}, \mathsf{config}_2)$ if every labelled object that satisfies the description $\underline{\mathsf{dgm}_1}(\mathsf{config}_1)$ also satisfies the description $\underline{\mathsf{dgm}_2}(\mathsf{config}_2)$, or, in other words, if

$$\underline{\mathsf{dgm}_1}(\mathsf{config}_1) \models \underline{\mathsf{dgm}_2}(\mathsf{config}_2)$$

is a theorem. For example, with harmless abuse of language, we may say that the labelled object $\underline{\mathsf{fred's\_front\_door}}(B, C, D, E)$ has the property $\underline{\mathsf{parallel}}(C - B, E - D)$.

We say that $\underline{\mathsf{dgm}_1}$ is_a $\underline{\mathsf{dgm}_2}$ if every configuration that realizes $\underline{\mathsf{dgm}_1}$ also realizes $\underline{\mathsf{dgm}_2}$ (or, equivalently, if, for every configuration config of appropriate type, $(\underline{\mathsf{dgm}_1}, \mathsf{config})$ has the property $(\underline{\mathsf{dgm}_2}, \mathsf{config})$); for example, according to our definitions, a $\underline{\mathsf{door}}$ is_a $\underline{\mathsf{wall}}$. This relation is a partial order on the set of all diagrams, and objects (being those diagrams that are realized by just one configuration) are precisely the minimal diagrams under this relation.

Example

The interplay of the different concepts that we have defined is exemplified by the equivalence (given the relevant diagram definitions) of the following statements:

- the configuration $(O, O + 3X, O + 3X + 7Y, O + 7Y)$ realizes the diagram $\underline{\mathsf{door}}$;
- for every variable configuration (A, B, C, D) in POINTS^4, the labelled object given by $\{A = O, B = O + 3X, C = O + 3X + 7Y, D = O + 7Y\}$ is a solution of the description $\{A - B + C - D = 0, B = A + 3X, C = B + 7Y\}$;
- $\{A = O, B = O + 3X, C = O + 3X + 7Y, D = O + 7Y\} \models \{A - B + C - D = 0, B = A + 3X, C = B + 7Y\}$;
- the assertion $(\underline{\mathsf{fred's\_front\_door}}(A, B, C, D), \underline{\mathsf{door}}(A, B, C, D))$ is a theorem;
- for every configuration $(\mathsf{exp}_1, \mathsf{exp}_2, \mathsf{exp}_3, \mathsf{exp}_4) \in \mathsf{POINTS}^4$, $\underline{\mathsf{fred's\_front\_door}}(\mathsf{exp}_1, \mathsf{exp}_2, \mathsf{exp}_3, \mathsf{exp}_4)$ has the property $\underline{\mathsf{door}}(\mathsf{exp}_1, \mathsf{exp}_2, \mathsf{exp}_3, \mathsf{exp}_4)$;
- $\underline{\mathsf{fred's\_front\_door}}$ is_a $\underline{\mathsf{door}}$. □

Quite generally, a configuration realizes a diagram $\underline{\mathsf{dgm}}$ if and only if the associated object is_a $\underline{\mathsf{dgm}}$.

8 Proof

The statement of an elementary assertion may be abbreviated (and its geometric meaning made more obvious) by allowing some of the equations to be replaced by diagram calls that produce them. Then elementary assertions have the form of Horn clauses with diagrams as explicit predicates and equations as implicit ones.

Thus there may be three different types of sentence involving predicates like "wall"—definitions, hypotheses (definition calls) and conclusions (or queries); and we also permit un-named equations as hypotheses or conclusion. Equations may be used directly, on a once off basis, as hypotheses of a theorem, or, for repeated use, may be incorporated in a data-base of diagrams. Current diagram definitions provide a global environment while un-named equations used as hypotheses provide a local one.

The simplest theorems are those with no hypotheses—they assert that a particular configuration realizes a diagram. A very basic mechanism available for proof is rewriting to normal form—for a single equation is valid if and only if its polynomial form is $0 = 0$. Rewriting alone gives, for example,

$$\begin{aligned}\mathsf{conc}\quad & \underline{\mathsf{parallelogram}}(O, O+X, O+X+Y, O+Y)\\ & \longmapsto (O+Y)-(O+X+Y)+(O+X)-O=0\\ & \longmapsto 0=0;\end{aligned}$$

in other words, the points $O, O+X, O+X+Y, O+Y$ form a parallelogram. Variable configurations give more interesting theorems:

$$\begin{aligned}\mathsf{conc}\quad & \underline{\mathsf{parallelogram}}(\tfrac{1}{2}(P+Q), Q, \tfrac{1}{2}(Q+S), \tfrac{1}{2}(P+S)\\ & \longmapsto \tfrac{1}{2}(P+Q)-Q+\tfrac{1}{2}(Q+S)-\tfrac{1}{2}(P+S)=0\\ & \longmapsto 0=0;\end{aligned}$$

in other words, for any points P, Q, S, the points $\frac{1}{2}(P+Q), Q, \frac{1}{2}(Q+S), \frac{1}{2}(P+S)$ form a parallelogram.

Another example:

$$\begin{aligned}\mathsf{conc}\quad & \underline{\mathsf{vertical}}(A+7V, A+7V+5Y)\\ & \longmapsto (A+7V+5Y)Y-(A+7V)Y=0\\ & \longmapsto 0=0;\end{aligned}$$

in other words, for any point A and vector V, the vector from $A+7V$ to $A+7V+5Y$ is vertical.

For our purposes a *proof* of an elementary assertion assert_1 may be taken to be a finite tree of elementary assertions with the following properties:

(1) the root is assert_1;

(2) every leaf has either a valid conclusion or unsatisfiable hypotheses (and hence is a theorem);

(3) every node is a theorem provided all its children are theorems.

In what follows we shall present some *inference rules*—they express how to determine legitimate children of a node. We shall sketch some sample proofs, but must leave the presentation of an efficient proof strategy for another occasion.

Rewriting of equations to normal form is, so to speak, a tacit inference rule (that reconciles our desire to work with polynomials with the practical need to represent them as expressions); we assume that it is automatically performed whenever new equations are created.

9 Reduction

An equation conc is said to be *reducible* via a description hyps if the head of some element of hyps occurs as a subproduct of some term of the conc; and then we say that conc *reduces to* conc′ *via* hyps if conc′ is the equation obtained from conc when some such occurence of the head of an equation is replaced by its body. In a proof, the node (hyps, conc) may be followed by (hyps, conc′); we call this the **reduction inference rule**. It is sound by virtue of the replacement property of equality.

Example

hyp $\quad \underline{\text{parallel}}(B-A, D-C) \longmapsto BD - BC - AD + AC = 0$
conc $\quad \underline{\text{coplanar}}(A, B, C, D) \longmapsto ABCD = 0$

Reduction replaces the conclusion by

$$-AC(BC + AD - AC) = 0$$

which rewrites to $0 = 0$, completing a proof. □

Note: in such **Examples** the rewrite symbol $\longmapsto$ and the equations that follow it are not part of the theorem statement (and do not represent user input) but are included simply to remind the reader of how the hypotheses and conclusion are rewritten when a proof attempt commences (and even, as with coplanar above, to implicitly supply a definition).

Example

hyp $\quad \underline{\text{door}}(A, B, C, D)$
$\qquad \longmapsto A - B + C - D = 0, B = A + 3X, C = B + 7Y$
conc $\quad \underline{\text{wall}}(A, B, C, D)$
$\qquad \longmapsto A - B + C - D = 0, (B - A)X = 0, (C - B)Y = 0$

This is the assertion that a door is_a wall, and is immediately provable by reduction. □

Example

hyp_1 $\underline{\mathsf{parallelogram}}(A,B,C,D) \longmapsto A - B + C - D = 0$
hyp_2 $\underline{\mathsf{midpoint}}(A,C,P) \longmapsto P = \frac{1}{2}(A+C) \longmapsto 2P - C - A = 0$
hyp_3 $\underline{\mathsf{midpoint}}(B,D,Q) \longmapsto Q = \frac{1}{2}(B+D) \longmapsto 2Q - D - B = 0$
conc $\underline{\mathsf{coincides}}(P,Q) \longmapsto P - Q = 0$

Proof by reduction is straightforward:

$$P - Q = \frac{1}{4}(A+C)(B+D) = \frac{1}{4}(A+C)(A+C) = 0 \ \square$$

For any finite set of equations hyps iterated reduction of any equation conc terminates after finitely many steps; however more than one irreducible equation may be obtainable. This may be remedied by critical pair completion. A variant of Buchberger's algorithm [3] can be used to replace the hypotheses by a semantically equivalent description via which reduction of any equation terminates uniquely. We refer to this process as *compilation*; though time-consuming, it need only be applied once to each diagram stored in the data-base.

Even after compilation, there may remain equations that are semantically implied by the hypotheses but cannot be deduced from them by reduction alone. Further inference rules are needed.

Example

Compiling the hypotheses

hyp_1 $AB = OX$
hyp_2 $BC = OY$

generates a new hypothesis

hyp_3 $OXC = OYA$

which may be used to deduce, by reduction alone, the equation

conc $OXYC = 0,$

but not, on the other hand, the equation

conc $ABC = OXY,$

which is also implied by the hypotheses. $\square$

10 Quantities

The expressiveness of Grassmann geometry springs, in part, from the fact that not only diagrams, but the quantities themselves may have geometrical meaning. For example $B - A$ is to be interpreted as the vector from A to B. Availability of both the notions of point and vector allows elementary theorems like the following to be succinctly and naturally expressed.

Example

def $\underline{\text{vector\_triangle}}(U, V, W) \longmapsto U + V + W = 0$
hyp_1 $\underline{\text{midpoint}}(B, C, D) \longmapsto D = \frac{1}{2}(B + C)$
hyp_2 $\underline{\text{midpoint}}(C, A, E) \longmapsto E = \frac{1}{2}(C + A)$
hyp_3 $\underline{\text{midpoint}}(A, B, F) \longmapsto F = \frac{1}{2}(A + B)$
conc $\underline{\text{vector\_triangle}}(D - A, E - B, C - F)$
$\longmapsto A + B + C - D - E - F = 0$

Proof by reduction is trivial. □

An attempt to prove an assertion with conclusion

$\text{target} = 0$

may be regarded more generally as seeking to compute the value of the target expression relative to the hypotheses. From this point of view a negative result may well convey useful information.

Example

hyp $\underline{\text{door}}(A, B, C, D)$
$\longmapsto A - B + C - D = 0, B = A + 3X, C = B + 7Y$
conc $AB + BC + CD + DA = 0$

The proof fails, concluding with the useful target value $42XY$—this shows that any door has area 21 (times the area of the "unit square"). There is a simple rationale for such interpretations, which we will not go into in the present paper. □

Example

If the diagonals of a quadrilateral (A, B, C, D) have midpoints M and N, and a pair of its opposite sides meet at P, then the triangle (P, M, N) has one quarter the area of (A, B, C, D). This is easily formalized as an elementary theorem:

hyp1 $\underline{\mathsf{collinear}}(A, D, P) \longmapsto AD + DP + PA = 0$
hyp2 $\underline{\mathsf{collinear}}(B, C, P) \longmapsto BC + CP + PA = 0$
hyp3 $\underline{\mathsf{midpoint}}(A, C, M) \longmapsto M = \frac{1}{2}(A + C)$
hyp4 $\underline{\mathsf{midpoint}}(B, D, N) \longmapsto N = \frac{1}{2}(B + D)$
conc $4(PM + MN + NP) = AB + BC + CD + DA$

And it is provable automatically by reduction:

$$\begin{aligned}
&4(PM + MN + NP)\\
&\quad = 2P(A + C) + (A + C)(B + D) + 2(B + D)P\\
&\quad = 2(AD + DP + PA) - 2(BC + CP + PB) + (AB + BC + CD + DA)\\
&\quad = AB + BC + CD + DA.
\end{aligned}$$

This is to be compared with the proof given in [4, page 55] in which two auxiliary points are created and two lemmas invoked. □

11 Instantiation

An expression of the form

$$\mathsf{num}_0 O + \mathsf{num}_1 X_1 + \cdots + \mathsf{num}_n X_n$$

has normal form 0 if and only if all the coefficients num_k are 0; in other words, the constants $O, X_1, \ldots, X_n$ are (linearly) independent over NUMBERS and hence form a basis for WORLD_n. A straightforward argument shows that ground points $\mathsf{pt}_1, \mathsf{pt}_2, \ldots, \mathsf{pt}_m$ (which necessarily belong to some finite-dimensional WORLD_n) are dependent if and only if

$$\mathsf{pt}_1 \mathsf{pt}_2 \ldots \mathsf{pt}_m = 0$$

A useful "incremental" variant of this is the following result.

Theorem

For ground points $\mathsf{pt}_1, \mathsf{pt}_2, \ldots, \mathsf{pt}_m$

$$\mathsf{pt}_1 \mathsf{pt}_2 \ldots \mathsf{pt}_m = 0$$

if and only if either

$$\mathsf{pt}_1 \mathsf{pt}_2 \ldots \mathsf{pt}_{m-1} = 0$$

or there exist scalars $\mathsf{num}_1, \mathsf{num}_2, \ldots, \mathsf{num}_{m-1}$ such that

$$\mathsf{pt}_m = \mathsf{num}_1 \mathsf{pt}_1 + \mathsf{num}_2 \mathsf{pt}_2 + \cdots + \mathsf{num}_{m-1} \mathsf{pt}_{m-1}$$

Proof: Suppose that $\mathsf{pt}_1\mathsf{pt}_2\ldots\mathsf{pt}_m = 0$. As observed above, there exist scalars $\mathsf{num}_1, \mathsf{num}_2, \ldots, \mathsf{num}_m$, not all zero, such that

$$\mathsf{num}_1\mathsf{pt}_1 + \mathsf{num}_2\mathsf{pt}_2 + \cdots + \mathsf{num}_m\mathsf{pt}_m = 0.$$

If num_m is non-zero then the second of the stated alternatives holds. Otherwise one of $\mathsf{pt}_1, \mathsf{pt}_2, \ldots, \mathsf{pt}_{m-1}$ is a linear combination of the rest and hence $\mathsf{pt}_1\mathsf{pt}_2\ldots\mathsf{pt}_{m-1} = 0$.

The converse is obvious. □

In the latter case of the theorem, the sum of the coefficients, $\mathsf{num}_1, \mathsf{num}_2, \ldots, \mathsf{num}_m$ is necessarily 1.

These facts concerning the WORLD (which are valid in all finite- dimensional Grassmann algebras) are not captured by the algebraic semantics presented so far, but can be incorporated by the device of introducing scalar variables and an appropriate new inference rule.

Instantiation inference rule

Suppose that an equation of the form

$$P_1P_2\ldots P_m = 0$$

may be deduced from an hypothesis of an assertion. Then the assertion may be replaced by a pair of assertions of which the first (called the *generic case*) is obtained by adding

$$P_k = a_1P_1 + a_2P_2 + \cdots + (1 - a_1 - a_2 - \cdots - a_{k-2})P_{k-1}$$

(where $a_1, a_2, \ldots, a_{k-2}$ are new scalar variables) to the hypotheses, and the second (the *exceptional case*) is obtained by adding

$$P_1P_2\ldots P_{k-1} = 0$$

to the hypotheses. □

This inference rule could be used to eliminate all hypotheses of degree greater than 1, but it is more efficient to confine its use to situations in which no reductions are possible.

The work of the prover may be considerably reduced, if desired, by making an assumption of genericity when appropriate: special relationships which are not implied by the data are assumed not to hold. This is analagous to the "negation as failure" rule.

In similar fashion, the fact that NUMBERS is a field is manifested by the **scalar division inference rule**: if an hypothesis of an assertion has the form

$$\mathsf{num.exp} = 0.$$

where num is a scalar expression and exp is an arbitrary expression, then the assertion may be replaced by a pair of assertions, of which the first (called the *generic case*) is obtained by replacing that hypothesis by

$$\mathsf{num} = 0,$$

and the second (called the *exceptional case*) by replacing it by

$$\mathsf{exp} = 0.$$

In particular, if exp is a non-zero ground polynomial, only the former case occurs.

Example

hyp $AM = MB$
conc $\underline{\text{midpoint}}(A, B, M) \longmapsto M = \frac{1}{2}(A + B)$

The hypothesis implies that $ABM = 0$. No reduction of the target being possible, M is instantiated as $aA + (1 - a)B$. Then hyp reduces to

$$(1 - 2a)AB = 0.$$

The assumption of genericity (namely that AB is nonzero) gives $a = \frac{1}{2}$ and hence $M = \frac{1}{2}(A + B)$. Finally, reduction of the target produces 0.

In this example, the assertion is also valid in the non-generic case as is easily proved:

hyp_1 $AB = 0$
hyp_2 $AM = MB$
conc $M = \frac{1}{2}(A + B)$

Instantiation produces, first, $B = A$ and, then, $M = A$, from which the conclusion follows immediately by reduction.

Since the converse (provable by reduction only) is also valid, the interpretation of the equation $AM = MB$ is forced. □

Example

hyp_1 $\underline{\text{vertical}}(A, B) \longmapsto BY - AY = 0$
hyp_2 $\underline{\text{parallel}}(B - A, D - C) \longmapsto BD - BC - AD + AC = 0$
conc $\underline{\text{vertical}}(C, D) \longmapsto CY - DY = 0$

Instantiating B as $A + bY$ reduces hyp_2 to

$$bDY - bCY = 0$$

from which the conclusion follows provided that b is not 0. If $b = 0$ the hypotheses are eliminated, and the assertion is not proved. □

12 Inconclusive proofs

An attempt to prove an assertion with satisfiable hypotheses succeeds if the final value computed for the target expression is 0 and fails if this final value is a non-zero constant. Otherwise the attempt is inconclusive; and this also may be useful.

Example

hyp $\underline{\text{parallelogram}}(A, B, C, D) \longmapsto A - B + C - D = 0$
conc $\underline{\text{collinear}}(A, C, D) \longmapsto AC + CD + DA = 0$

Although the attempt to prove this is inconclusive it terminates with the useful target value $AB + BC + CA$. This reveals that the assertion would become valid either if

$$AB + BC + CA = 0$$

was added to the hypotheses (that is if A, B and C were collinear), or if the conclusion was replaced by

$$AC + CD + DA = AB + BC + CA$$

(that is by the conclusion that the triangles (A, C, D) and (A, B, C) have equal areas). Note that if with the same hypothesis an attempt is made to prove that $AB + BC + CA = 0$ the returned target value is $AB + BC + CA$, which reveals nothing; the user is able to exercise some degree of control by the ordering he sets up (by the names he gives them) for the points of the assertion. □

Example

hyp_1 $Q = O + X$
hyp_2 $R = O + 3X + 2Y$
hyp_3 $P = O + xX + yY$
conc $\underline{\text{collinear}}(P, Q, R) \longmapsto PQ + QR + RP = 0$

We must compute the value of

$$\text{target} = PQ + QR + RP.$$

Reduction gives

$$\text{target} = 2(y - x + 1)XY.$$

In the generic case the conclusion is false. In the exceptional case, where

$$y - x + 1 = 0$$

the conclusion holds. Thus the equation of the line through P and R is obtained. □

13 Coordinate proofs

The following inference rule may be used to replace any assertion by an equivalent assertion of standard polynomial algebra over the rationals.

Coordinatization inference rule

An equation of an assertion that has the normal form

$$c_1\mathsf{exp}_1 + \cdots c_n\mathsf{exp}_n = 0$$

where $\mathsf{exp}_1, \ldots, \mathsf{exp}_n$ are ground product terms and $c_1, \ldots, c_n$ are scalars may be replaced by the set of equations

$$c_1 = 0, \ldots, c_n = 0. \ \square$$

The soundness of this rule follows from the fact that each exp_i must be a product of distinct elements of the sequence $O, X_1, X_2, \ldots$ and that any finite set of distinct such products is necessarily independent.

To prove a theorem using coordinates is effectively to prove it at the semantic level. For example the coordinate proof below shows directly that any object labelling that satisfies the hypotheses of the assertion satisfies its conclusion.

Example

The elementary assertion

hyp_1 $D = \frac{1}{2}(A + B)$
hyp_2 $E = \frac{1}{2}(A + C)$
conc $4ADE = ABC$

is easily proved by reduction:

$$\begin{aligned} 4ADE - ABC &= A(A + B)(A + C) - ABC \\ &= 0. \end{aligned}$$

We may produce a coordinate proof by incorporating the following equations as additional hypotheses:

$$\begin{aligned} A &= O + a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 \\ B &= O + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 \\ C &= O + c_1X_1 + c_2X_2 + c_3X_3 + c_4X_4 \\ D &= O + d_1X_1 + d_2X_2 + d_3X_3 + d_4X_4 \\ E &= O + e_1X_1 + e_2X_2 + e_3X_3 + e_4X_4 \end{aligned}$$

Pre-processing of the hypotheses followed by application of the coordinatization inference rule to each of them yields the coordinate constraints:

$$2d_i - a_i - b_i = 0, \quad 1 \leq i \leq 4,$$
$$2e_i - a_i - c_i = 0, \quad 1 \leq i \leq 4.$$

And the conclusion, as one easily calculates, is equivalent to the scalar equations

$$4(d_ie_j - d_je_i - a_ie_j + a_je_i + a_id_j - a_jd_i)$$
$$-(b_ic_j - b_jc_i - a_ic_j + a_jc_i + a_ib_j - a_jb_i) = 0, \quad 1 \leq j < i \leq 4.$$

We see that proving this theorem is equivalent to a standard problem of inferring polynomial equations (6 of them, each homogeneous of degree 2) from polynomial equations (8 of them, each of degree 1) in several variables (20 of them) and could be solved, though with much unnecessary labour, by the Grobner basis method [3]. In practice one may well be happy with proofs that are valid at least in three dimensions; in that case only three basis vectors are needed and the complexity of the equivalent set of scalar equations is reduced accordingly. As a matter of fact, this example is necessarily planar (because the hypotheses entail that $ABCD = 0$ and $ABCE = 0$) and hence, if one could first somehow realize this fact, a genuine (not dimension-restricted) proof could be obtained by deriving just one second degree equation from 4 first degree equations in 10 scalar variables. □

In two dimensions, coordinate proofs become highly feasible as we know from the remarkable China prover (see [20], [5], and other papers of these authors). The point to be noted is that by going to coordinates Grassmann geometry computations are reduced to standard polynomial algebra; hence, what is known about the power of such methods is pertinent to Grassmann geometry. At the same time, the comparison makes evident the relative efficiency, even for such a simple example as the one above, of a proof method that handles geometric constraints directly without going to coordinates.

14 Conclusion

The specification of a diagram, as described here, is geometrical in character. To produce a picture, topological information must also be provided. We shall present elsewhere an extension of the formalism which allows this to be done in an elegant way; moreover topological properties can be proved by equational reasoning.

A limitation of the language as outlined is that it is confined to affine geometry. This seemed a reasonable discipline to impose in a paper whose purpose is to introduce the idea of a diagram. But the point may also be made that affine geometry is fundamental to applications such as computer- aided architectural design, and that it is in the nature of things that affine geometry can be more efficiently implemented than its extensions. In fact, though there is much work to be done, we believe that the Grassmann geometry formalism can be generalized to take in convex geometry, on the one hand, and metric geometries, on

the other. But the idea of a diagram is independent of particular formalisms, and indeed is pertinent to areas other than geometry.

The pencil sketch which a mathematician draws when he is trying to prove a theorem about parallelograms is not, of course, a generic parallelogram, only a helpful representation of one. What, one may ask, is it that is so represented? The ideal parallelogram does not share all the properties of any given parallelogram, yet has the properties that are shared by all particular parallelograms. It might be thought that the set of all configurations that "are" parallelograms adequately represents the idea of a parallelogram, but this is not enough for practical purposes; for in order to speak about properties, points must be given names. At the same time, the properties themselves must be independent of the names—there must be names, but what they are does not matter. It is this conception that the notion of a diagram attempts to capture.

Acknowledgements

The author thanks Alan Robinson and Tim Stokes for conversations pertinent to the paper, and José Meseguer for drawing attention to several items in the bibliography.
Note: the bibliography, though not comprehensive, includes a representative selection of recent papers relevant to geometry theorem proving, and, in particular, reports of the current effort to integrate algebraic reasoning with logic programming.

References

[1] Farhad Arbab and Jeanette M.Wing, "Geometric reasoning: a new paradigm for processing geometric informatiom", *IFIP-86,* 107–121.

[2] Alan Borning, "The programming language aspects of ThingLab, a constraint-oriented simulation laboratory", *ACM Trans. Prog. Languages and Systems* **3** (1981), 353–387.

[3] Bruno Buchberger, "Grobner bases: an algorithmic method in polynomial ideal theory", 184–232 . in *Multi-dimensional Systems Theory,* ed. R. K. Bose, Reidel, 1985.

[4] H.S.M.Coxeter and S.L.Greitzer, *Geometry Revisited,* MAA New Mathematical Library, 1967.

[5] Shang-Ching Chou, "Proving elementary geometry theorems using Wu's algorithm", 243–286, in *Automated Theorem Proving: After 25 Years* , ed. W. W. Bledsoe and D. W. Loveland, American Mathematical Society , 1984.

[6] Shang-Ching Chou and William F. Schelter, "Proving geometry theorems with rewrite rules", *J. Automated Reasoning,* **2** (1986),153–273.

[7] Shang-Ching Chou, William F. Schelter and Jin-Gen Yang, "Characteristic sets and Grobner bases in geometry theorem proving", this Colloquium.

[8] Helder Coelho and Luiz Moniz Pereira, "Automated reasoning in geometry theorem proving with Prolog", *J. Automated Reasoning,* **2** (1986), 329–390.

[9] Mehmet Dincbas, Helmut Simonis and Pascal van Hentenryck, "Extending equation solving and constraint handling in logic programming", this Colloquium.

[10] Desmond Fearnley-Sander, "Affine geometry and exterior algebra", *Houston J. Math.,* **6** (1980), 53–58.

[11] Desmond Fearnley-Sander, "Using and computing symmetry in geometry proofs", *University of Edinburgh, Department of Artificial Intelligence Research Paper* **257**, 1985.

[12] Desmond Fearnley-Sander, "Diagrams in Grassmann geometry", *University of Tasmania, Department of Mathematics, Technical Report,* 1986.

[13] Joseph A. Goguen, "Modular algebraic specification of some basic geometrical constructions", to appear.

[14] C. M. Hoffmann and M. J. O'Donnell, "Programming with equations", *ACM Trans. Prog. Lang. and Sys.,* (1982), 83 112.

[15] G.Huet and D.C.Oppen, "Equations and rewrite rules", in *Formal Languages: Perspectives and Open Problems,* ed R.V. Book, Academic Press, 1980.

[16] Joxan Jaffar and Jean-Louis Lassez, "Constraint logic programming", *Proc. ACM POPL Conference* (1987).

[17] Deepak Kapur, "Using Grobner bases to reason about geometry problems", *J. Symb. Comp.,* **2** (1986), 399–208.

[18] B.Kutzler and S.Stifter "On the application of Buchberger's algorithm to automated geometry theorem proving", *J. Symb. Comp.,* **2** (1986), 389–397.

[19] Christopher J. van Wyck, "A high-level language for specifying pictures", *ACM Trans. Graphics,* **1** (1982), 163–182.

[20] Wu Wen-Tsun 86, "Basic principles of mechanical theorem-proving in geometries", *J. Automated Reasoning,* **2** (1986), 221–252.