

Main problems of diagrammatic reasoning.

Part I: The generalization problem

Zenon Kulpa

*Institute of Fundamental Technological Research of the Polish Academy of Sciences
ul. Świętokrzyska 21, 00-049 Warsaw, Poland*

Abstract. The paper attempts to analyze in some detail the main problems encountered in reasoning using diagrams, which may cause errors in reasoning, produce doubts concerning the reliability of diagrams, and impressions that diagrammatic reasoning lacks the rigour necessary for mathematical reasoning. The paper first argues that such impressions come from long neglect which led to a lack of well-developed, properly tested and reliable reasoning methods, as contrasted with the amount of work generations of mathematicians expended on refining the methods of reasoning with formulae and predicate calculus. Next, two main groups of problems occurring in diagrammatic reasoning are introduced. The second group, called diagram imprecision, is then briefly summarized, its detailed analysis being postponed to another paper. The first group, called collectively the generalization problem, is analyzed in detail in the rest of the paper. The nature and causes of the problems from this group are explained, methods of detecting the potentially harmful occurrences of these problems are discussed, and remedies for possible errors they may cause are proposed. Some of the methods are adapted from similar methods used in reasoning with formulae, several other problems constitute new, specifically diagrammatic ways of reliable reasoning.

Keywords: diagrammatics, diagrammatic reasoning, rigorous reasoning, reasoning errors, generalization, diagram particularity, divergence.

1. Introduction

Diagrammatics is a new discipline of research¹ devoted to the investigation of *diagrams* as a means for representation and processing of knowledge. Knowledge processing, usually called *reasoning*,² can be, and often is, used for argumentation in many areas. If the subject matter is mathematics, it is required that the reasoning fulfills strict standards of objectivity, rigour, and reliability in order to be considered valid and dependable. Centuries of work of mathematicians and philosophers resulted in the development of ways of obtaining such rigour and

¹ The “official” beginning of the discipline is usually assumed to be the first workshop on the subject: The AAAI Spring Symposium on *Reasoning with Diagrammatic Representations* (DIAGRAMS, 1992).

² Whether there exists some sort of knowledge processing which is not a reasoning will not bother us here.



reliability with propositional knowledge representations, mostly mathematical formulae. The final conclusion of this process occurred around the break of XIX and XX centuries, when the invention of *predicate calculus* by Frege (1879 [1967]) and works of a number of other mathematicians, especially Hilbert, resulted in the notion of *formal* reasoning (usually based on the language of predicate calculus) which can be fully objective and rigorous, as consisting of purely mechanical, finitistic symbol rewriting. However, this development has its drawbacks too.

First, the strictly formal techniques are rarely actually used by mathematicians, because they are tedious and time consuming, produce very long sequences of illegible formulae, and, somewhat paradoxically, are very unreliable — people are not very good mechanical rewriters of large masses of symbols and tend to make many errors in the process. The advent of computers raised much hope, as they are very good in fast, blind and reliable symbol crunching. The hopes were not fulfilled, though — it was realized that the crucial thing here is how to find a way through the maze of possible rewritings so that they are laid along the route towards our desired goal, say, the thesis we are arguing for. Mathematicians find this route using understanding of the meaning of formal symbol sequences and that archenemy of formal reasoning advocates — mathematical intuition. This proved very hard to teach computers to do, and without intuitive guidance the rewriting process for any non-trivial problem suffers a combinatorial exponential explosion in the number of possible rewritings, which even the fastest computers cannot handle.

Therefore, automatic theorem proving did not achieve much success — the computers were more successfully used as verifiers of proofs made by humans and formalized in languages readable by both humans and machines (Muzalewski, 1993). This also brought attention to the fact that formal reasoning is only a part, and often a small part, of a mathematician's work. As Poincaré observed “It is by logic that we prove, but by intuition that we discover.” And there are not many ready-made tools for helping mathematicians in their informal work, useful not only in the context of discovery, by the way. It is not surprising if we consider the amount of work expended by the formalization movement just to get rid of unreliable intuition and replace it by purely mechanical reasoning. But as often happens in such cases, the proper approach does not involve *replacing* one way by another, but instead *augmenting* the one by the other. And indeed, mathematicians usually produce informal proofs using much intuition and informal leaps of imagination, but still maintaining a certain discipline and rigour that convinces them that the result in principle *can* be formalized if need be. However, it is hard to hear a convincing answer to the question what exactly makes them so

sure of that possibility (it seems that intuition still plays an important role here).

In the process of formalizing mathematics, an old and venerable tool of informal reasoning has been ridiculed and dropped. It was the diagram. It was used in mathematics since times immemorial, and its use was widespread, from geometrical diagrams of Euclid to the Argand diagram and its derivatives in complex analysis (Needham, 1997). Diagrams were, however, accused of being too difficult to use, unreliable and error-prone, and, of course, unfit for formalization. It is true that their nature and use differ considerably from the classic representation and reasoning using formulae. However, taking into account that until recent times no serious attempts have been made to understand and tame this reasoning tool, compared with the huge amount of work expended on understanding and making rigorous propositional reasoning, it is hardly surprising that diagrams seem difficult, unreliable, and not rigorous enough.

Concerning the formalization of diagrams, certain simple kinds, like Venn diagrams (Shin, 1994), dot-pattern arithmetic diagrams (Jamnik, 2001), or elementary geometry diagrams (Luengo, 1995; Miller, 2001) have already been formalized, though the results are not quite satisfactory, especially from the pragmatic point of view. But at least these attempts demonstrated that it is possible in principle. It seems that the attempts of satisfactory formalization of diagrammatic reasoning systems will cause certain important changes in our understanding of the notion “formal reasoning.” This interesting issue is, however, outside the scope of this paper. Moreover, formalization of diagrammatic reasoning is not an ultimate goal of diagrammatic research, as we rarely use strictly formal methods in practice, no matter whether we use diagrams or formulae. The possibility of formalization usually serves only as a sort of certificate that the given reasoning method is valid and rigorous enough to be dependable.

Because of the essential differences between classic representations and reasoning using formulae, and diagrammatic representations and reasoning, especially in mathematics, various techniques developed for the former do not always transfer directly and naturally to the latter. That causes some troubles, as diagrammatic techniques are not yet well developed, due to the long neglect of this representation and reasoning approach. These troubles are usually perceived as “problems with” or “drawbacks of” diagrammatic reasoning, and are considered as indications of a generic inferiority of diagrams as representation and reasoning tools. Many such problems are mentioned in the literature, under various names. One of the more comprehensive lists is given in (Winterstein, 2004). Under closer scrutiny, however, many of them can be attributed

to an inadequate visual (diagrammatic) language used to represent the domain of interest. Extending the language a little causes many of these problems to disappear. Other problems can be recognized as caused by a few generic features of diagrams, which require the development of new, specifically diagrammatic approaches to reasoning.

Diagrams, due to their two-dimensionality and a rich set of graphical properties and relations, allow for building much richer, more complex, and more efficient meaningful structures than one-dimensional symbol strings, and exhibit a number of phenomena unknown and not occurring in the world of formulae. Users of formulae had already tried to tap some of these possibilities, though with reservations: formulae expanded into the second dimension, and acquired certain diagrammatic elements too. Consider multilevel fractions (including infinite chain fractions), intricate multilevel subscript and superscript structures, matrices with various block notations (Pollet et al., 2004), proliferation of graphically complex symbols (Roman letters of several styles like gothic, calligraphic, etc., Greek and Hebrew letters, graphical symbols of operators, roots, integrals, etc.), abstraction devices like iterated summing and multiplication symbols Σ and Π , ellipsis (Foo et al., 1999; Bundy and Richardson, 1999; Pollet et al., 2004), arrow structures of category theory, and so on. And all that more than a hundred years since it was shown that it should suffice to use only letters and digits, two quantifier symbols, a few logical operators, some punctuation marks and parentheses, and everything significant in mathematics can be written down with the simple syntax of predicate calculus.³ May be it is possible, but then why not go even further and use only two symbols in formulae, say 0 and 1. This will also suffice — in fact, a swiftly growing amount of knowledge and information, including sounds and pictures, is indeed becoming encoded in this way... But people know better, leaving this level of encoding to machines, while using much richer systems of notations, including diagrams, at their human level.

³ It is interesting that Frege (1879 [1967]), the founder of predicate calculus, should propose instead a system of quite intricate *diagrammatic notation* for his new logical calculus. His diagrams did not gain any popularity and were forgotten, especially after Peano 15 years later introduced a textual notation for predicate calculus similar to that we use today. According to Greaves (2001), the contemporaries of Frege complained that his diagrams were hard to understand and use, especially due to problems with typesetting and printing! Frege himself retorted with “comfort of the typesetter is certainly not the *summum bonum*,” but without success. Today, with the help of computer typesetting, the situation might be quite different. May be it is the high time for diagrammatic researchers to revive Frege’s *Begriffsschrift* rather than to develop yet another version of formalized Venn diagrams. Indeed, some work seems to have started in this direction, see (MacInnis et al., 2003).

The aim of this paper is to show that, despite many criticisms to the contrary, diagrammatic reasoning can be made equally reliable and dependable as formulae, provided we expend some effort on systematization of sources of diagrammatic errors and finding remedies for them. This does not mean formalization, though it does not exclude it, as a means to certify (in a sense) the reliability of the tool.

In the next section of the paper two main groups of problems with diagrammatic reasoning, as advanced in the literature, are introduced. They are called, collectively, the *generalization* problem (called also the *universal quantification* problem) and the *imprecision* problem (or *existential quantification* problem). The second one is more difficult, as it has no practically important counterpart in reasoning with formulae, and is only summarized briefly in the next subsection. Its more detailed analysis is postponed to another paper (Kulpa, 2009a, forthcoming).

The next section is devoted entirely to the analysis of the first, easier problem, which partially occurs with formulae as well. In its subsections all the main aspects (particular sub-problems) of that problem are analyzed, concerning their nature and causes, the methods of detecting the occurrences of the problems are discussed and remedies for possible errors they may cause are proposed. Some of the methods are adaptations of that used in reasoning with formulae, the others are based on new, specifically diagrammatic approaches. In the conclusions section the results achieved by the analysis are evaluated and directions of further investigation are outlined.

2. Two main problem classes

After a closer look, many of the seemingly disparate “problems” with diagrammatic reasoning can be grouped into two main clusters:

- (I) The universal quantification problem (or \forall -*problem* for short):
Also called the **generalization problem**, *diagram particularity* problem, *representation of variables* problem, or the problem of *representation of quantifiers*.
- (II) The existential quantification problem (or \exists -*problem* for short):
Also called the **diagram imprecision problem**, *diagram perception* problem, *impossible cases* problem, or *limited expressiveness* problem.

The problems from the first of these groups seem easier, as some of them occur in similar form in formulae too. These problems will be analyzed in more detail in the rest of the paper. The second problem group is specific to diagrams and harder to resolve. It is only briefly

summarized below. Its more detailed analysis is the subject of (Kulpa, 2009a, forthcoming), the second of this series of papers.

2.1. THE \exists -PROBLEM

A common problem for *all* kinds of representations used for mathematical reasoning is how to ensure that the mathematical object we want to reason about actually exists. It is often claimed that diagrammatic representations have solved this problem, because they have a nice property of *self-consistency* (Lemon and Pratt, 1997; Kulpa, 2003a). There is no trouble with lying in a language or with a mathematical formula: “This black cat is white,” or “ $2 \times 2 = 5$.” However, it is impossible to draw a black cat which is simultaneously white, or arrange some tokens in two rows and two columns so that the number of needed tokens will be other than 4, see Fig. 1.

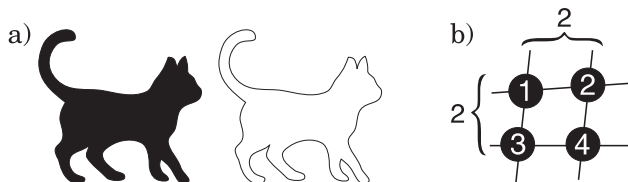


Figure 1. Self-consistency of diagrams: the black cat cannot be white (a), 2 (rows) \times 2 (columns) = 4 (tokens) always (b).

This property of diagrams is often summarized by the saying:
“If it can be drawn, it exists.”

It is also sometimes formulated as:

“If it cannot be drawn, it does not exist.”

However, this is true only for a limited class of discrete diagrams (where only a positioning combinatorics of discrete tokens matters), such as logical tables (say, *Karnaugh maps*), or other simple diagrams of structural type, like *Venn diagrams* (Shin, 1994). It is not so with more complex diagrams, especially “metric” diagrams representing some continuous domains, like geometry. One reason is a limited accuracy of rendering and perception of physical diagrams, which leads to the situation where two qualitatively different domain configurations result in identical (physically or perceptually) physical diagrams.

E.g., there is no way to distinguish diagrammatically between the $[0, 1]$ segment of a number line which contains all real points and its subset containing only the rational points (Winterstein, 2004). Yet another simple example — if three non-collinear points are, for a given degree of accuracy, sufficiently near to collinearity, their physical rendering will be indistinguishable from the configuration of three truly collinear points.

It should be noted that imprecision as such does not necessarily lead to reasoning errors. There are many cases where very sketchy and imprecise diagrams lead to rigorous and valid reasoning. This is so because we do not reason with the physical diagram itself, but with its idealization in our minds, where lines become perfectly straight, right angles perfectly right, etc. It can be interpreted as a specifically diagrammatic kind of symbolization (as with formulae, where the letter “ x ” means the same ideal symbol no matter in what font it was printed or in how peculiar a way it was handwritten, within certain recognition limits). Imprecision errors occur when such idealization goes against the intentions of the diagram creator. Consider the three points mentioned above — assume that in our diagram they should be non-collinear, but are placed in a diagram in such positions that the idealization process can easily mark them as collinear. If the conclusion of our reasoning crucially depends on the collinearity (or not) of these points, we will inevitably fall into a reasoning error.

This property of diagrams can produce geometric configurations that look very convincing despite the fact that they cannot exist (so-called “impossible cases”), like the diagram used for one of the cases of the fallacious “proof” that all triangles are isosceles (Dubnow, 1955 [1963]; Maxwell, 1959), see Fig. 2(a). This kind of imprecision is due to physical limitations of the diagrammatic medium, and of the processes of drawing a diagram and its perception. In the world of formulae it is practically insignificant, as it occurs only in rare cases like possible confusion of the letter “O” with the digit “0,” or the letter “l” with the digit “1,” and can be easily amended.

It should be noted that there are situations where such impossible diagrams are actually useful, consider e.g. the *reductio ad absurdum* proofs (probably the first example of the use of such an impossible diagram is Euclid I.6). The important differences between the harmful appearances of impossible cases and useful derivation of them are postponed to another paper (Kulpa, 2009a, forthcoming).

However, there is another, specifically diagrammatic limit — the limited expressiveness of the two-dimensional plane. It manifests itself in such effects as Helly’s Theorem (Lemon and Pratt, 1997) and representation of many-dimensional spaces. For example, rendering of three-dimensional bodies onto a plane causes severe loss of information, allowing for the drawing of so-called “impossible figures” (Kulpa, 1983; Kulpa, 1987; Kulpa, 2003a), see Fig. 2(b).

There are various ways to detect such situations and avoid possible errors in reasoning caused by them. These ways, as well as deeper analysis of the imprecision problem, are the subject of another paper (Kulpa, 2009a, forthcoming).

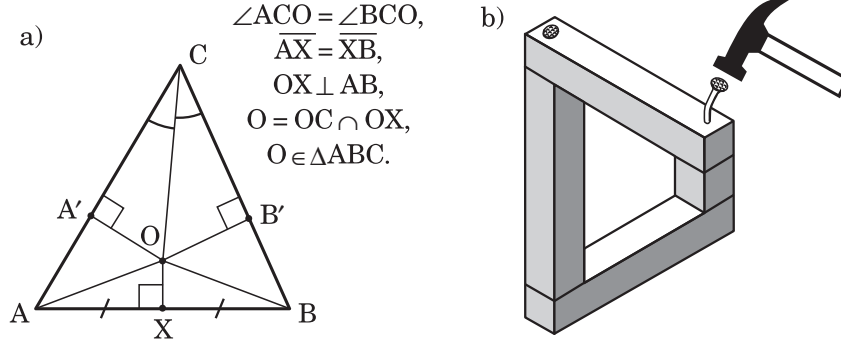


Figure 2. An impossible geometrical configuration (a), and the construction of an impossible quadrilateral (b).

3. The \forall -problem

A simple theorem is a statement that a certain object X has a certain property R , that is, $R(X)$ in predicate calculus form. That may seem a small piece of information, but what if the object X is a set, possibly infinite, and the property of that set says that all of its elements have some other property? The statement of such a more informative theorem will look like: $R(X) = (\forall x \in X)P(x)$. It states that all objects from a certain set X have the (new) property P . We can see three essential elements of this statement:

- The *variable* x , that is an object that could take *any value* from some set (here X).
- The *universal quantifier* \forall which states that the subsequent property holds for *all values* of the variable x taken from the set X .
- The *set* X of values the variable x can take.

The proof of the theorem involves certain reasoning (whatever that means), which demonstrates convincingly that $P(x)$. Well, how then can we be sure, that this statement can be generalized to all x in the set X ? The statement of the theorem only states the resulting claimed fact: the mere use of the quantifier \forall does not ensure its universal truth. The answer comes from consideration of the structure of the said reasoning. Reasoning must start from some assumptions. As the goal is to demonstrate some property of x , the assumptions must consist of some other properties of x , let us call them collectively $Q(x)$. Now if we are sure that these other properties of x hold for all x in X , then we are sure that $P(x)$ also holds for all objects x in X . Using formulae: $((\forall x \in X)Q(x) \ \& \ Q(x) \Rightarrow P(x)) \Rightarrow (\forall x \in X)P(x)$.

Certain “problems with diagrams,” usually discussed separately, after some scrutiny reveal that they are different aspects of the same general problem of performing with diagrams the proof scheme outlined above. The principal aspects are as follows:

- (a) Generalization, see Sections 3.1 and 3.2.
- (b) Particularity, see Sections 3.3 and 3.4.
- (c) Variable representation, see Sections 3.5 and 3.6.
- (d) Quantifier representation, see Section 3.7.
- (e) Representation of sets, see Section 3.8.

Let us analyze them in turn.

3.1. DIAGRAMMATIC GENERALIZATION

The diagrammatic generalization problem is usually stated in this way:

After we have proved a theorem using a diagram, how can we legitimately generalize the configuration of this diagram to a wide (usually infinite) class of configurations, and to what class exactly?

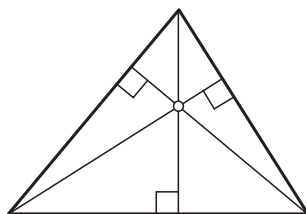


Figure 3. Altitudes in a triangle.

Let us illustrate the problem with a simple example. Consider the triangle shown in Fig. 3. The three altitudes of the triangle intersect at a single point inside the triangle. A more or less diagrammatic proof can be conducted to demonstrate validity of this observation (we are omitting it for brevity). Now we are tempted to generalize that property of this particular triangle to all triangles:

“Altitudes of *all triangles* intersect at a *single point inside a triangle*.”

However, after drawing another triangle (Fig. 4(a)) and repeating the construction, we see that the point of intersection now lies *outside* the triangle. It could also lie at the triangle’s vertex (when the triangle is right, Fig. 4(b)). Thus, our initial generalization was wrong.

There are several ways out of this predicament. Namely, we may either narrow the set X , such that all the elements of this smaller set have the obtained property:

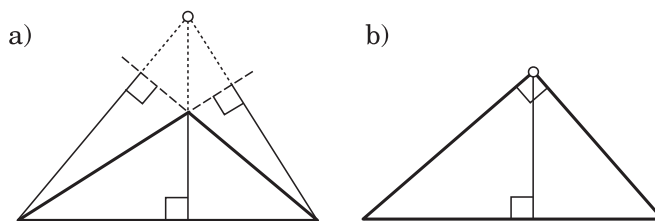


Figure 4. Altitudes in obtuse (a) and right (b) triangles.

“Altitudes of all *acute triangles* intersect at a single point inside a triangle.”

or instead restrict the property we are proving:

“Altitudes of all triangles intersect *at a single point.*”

We can also combine these possibilities, obtaining as a result three separate theorems for the three possible cases:

“Altitudes of all triangles intersect *at a single point*, which lies *inside* the triangle in *acute triangles*, *outside* the triangle in *obtuse ones*, and at a *vertex* in *right triangles.*”

3.2. GENERALIZING FROM DIAGRAMMS SAFELY: DIVERGENCE RULE

To solve the generalization problem as formulated in Section 3.1, we must first recognize what variables are involved (see Section 3.5), and then what sets of values they can range over. Such sets usually divide into natural subsets of “similar” objects, while objects from different subsets have different “structures,” indicated e.g. by the fact that their diagrammatic representations differ significantly. Thus, the construction of our diagrammatic proof becomes different for different subsets, as it needs different diagrams for every subset, possibly leading to different conclusions, as in the triangle example above. When the set in question is homogenous in this respect, our construction should be valid for all objects in the set, so that there is nothing more to do, as no wrong generalization may occur. Otherwise, we should conduct the proof separately for every case. That is summarized by the method of *divergence* (Kulpa, 2003b):

The divergence rule. *Check whether the set X divides into structurally different subsets requiring significantly different diagram configurations to represent reasoning about properties of their elements. If so, diverge the reasoning into separate cases, checking whether these different configurations produce the same property of the object x as that proved for other cases. If it is true for all subsets, the general theorem is proved for all $x \in X$, otherwise either:*

- *the thesis must be restricted to apply to some appropriate subset of X , or*
- *the property must be restricted to such a common part of the properties demonstrated for all the cases which is valid for all subsets (i.e., for the whole set X), or otherwise*
- *the theorem should be split into a number of different theorems valid for separate subsets of the set X .*

Problems can occur when there is an infinite number of such subsets of X . Fortunately, it is usual in such cases that even when structures of the subsets are different, they nevertheless exhibit a pattern of systematic change of structure which can be used, with an appropriate form of mathematical induction, to prove the thesis for all subsets on the basis of a finite number of proofs for several particular cases. This technique is applied e.g. by Jamnik (2001) to prove diagrammatically certain theorems of number theory. An example of such a theorem is the Nicomachus diagram discussed in Section 3.8.

Of course, finding an appropriate division of the set or even only spotting the necessity of such a division may be hard. The divergence rule makes us more aware of such necessity and explains how the division should be properly done. It is especially important in diagrammatic reasoning, where such case-splitting seems to be more common than in reasoning with formulae. Diagrams also help in spotting such situations: after drawing a diagram to prove something about, say, triangles, it is rather easy to see that our diagrammatic construction may look quite differently for a different kind of a triangle, as in the example above. Additional help comes with explicit representations of sets in diagrams, see Section 3.8, or using interactively transformable dynamic diagrams (King and Schattschneider, 1997; Le and Kulpa, 2003; Le and Kulpa, 2004). Note that the division of X into subsets is usually different for different properties we want to prove about the set. Compare the Altitudes in a Triangle theorem, where we were forced to divide the set of triangles into three subsets (distinguishing isosceles triangles was not necessary) with Arnheim’s proof of the sum of angles in a triangle (see Fig. 8), where no division of this set is necessary.

As concerns the Altitudes in a Triangle theorem of Section 3.1, the divergence rule leads to the following procedure. First, the variable x denotes seemingly a triangle, and second, the set X over which it varies is the set of all triangles. Does this set fall apart into subsets that have different structures with respect to the positions of altitudes, whose properties we attempt to prove? Indeed it does — in some kinds of triangles some altitudes lie *outside* them, a difference from our first

diagram in Fig. 3. Thus, the different subsets here are the sets of *acute* and *obtuse* triangles, and we should add also the intermediate case of *right* triangles for safety. Diverging our diagrammatic proofs to these three cases we obtain the three diagrams of Fig. 3 and 4. Then, as the properties of the altitude intersection point are different for each case, we can, as shown in Section 3.1:

- Restrict the generalization to an appropriate subset of X : “... all acute triangles ...”
- Restrict the thesis to the common part of all sub-theses: “ ... at a single point.”
- Divide the theorem into three theorems for separate cases.

3.3. DIAGRAM PARTICULARITY

Another formulation of the generalization problem is called *particularity* of diagrams. Considering the subformula $x \in X$, generality concentrates on finding the proper set X and analysing its structure, while particularity starts from the object x depicted in the diagram. In most of the diagrammatic literature it is claimed that:

Diagrams suffer from the property of *particularity*, namely, a diagram can actually prove the desired property only for some *particular object* x_0 of the set X , which is depicted in the diagram.⁴

Thus, we are again confronted with the generalization problem — to obtain the required proof for all objects in X we should properly *generalize* from this particular case to all elements of X — a task which is allegedly hard and error-prone with diagrams.

Again, this seems plausible enough. As a consequence, one may insist, to prove something diagrammatically, one would have to provide a number (usually infinite) of diagrammatic proofs for every $x \in X$. That is of course prohibitive, so that general and rigorous diagrammatic proofs seem to be impossible.

However, let us again see how this is done with formulae. After all, in proving with formulae we do not repeat the proof an infinite number of times for all members of some infinite set X . The direct analogue of the above formulation of diagram particularity would be an attempt to “prove” the formula for the square of a binomial in the following way. Start from, say, $(1+3)^2$ and proceed according to the rules of arithmetic: $(1 + 3)^2 = 4^2 = 16$. Then take $1^2 + 2(1 \cdot 3) + 3^2 = 1 + 6 + 9 = 16$. We

⁴ This property of diagrams was mentioned already by Aristotle in *On Memory and Reminiscence*.

obtained the same result, hence $(1 + 3)^2 = 1^2 + 2(1 \cdot 3) + 3^2$. Now we confront the same problem as that with the diagrammatic reasoning procedure outlined above: how to generalize that particular result to the general rule $(a + b)^2 = a^2 + 2ab + b^2$? It seems hopeless, as with diagrams...

Fortunately, with formulae we usually take quite another way to prove the theorem. Instead of using particular numbers, we write the formula in letters: $(a+b)^2$, and pretending the letters represent some unspecified numbers, we perform arithmetic operations on them according to the rules of arithmetic reasoning (tautologies of real algebra):

$$\begin{aligned}
 (a + b)^2 &= (a + b)(a + b) && \text{– by the definition} \\
 & && \text{of a square power,} \\
 &= aa + ab + ba + bb && \text{– multiplication of sums rule,} \\
 &= a^2 + ab + ab + b^2 && \text{– squaring and commutativity} \\
 & && \text{of multiplication,} \\
 &= a^2 + 2ab + b^2 && \text{– replacing multiple addition} \\
 & && \text{by multiplication.}
 \end{aligned}$$

Because all of these transformations are independent of the concrete values of the numbers that can be put at the place of letters a and b , the resulting equality $(a + b)^2 = a^2 + 2ab + b^2$ is valid for *all* numbers a and b , that is, we can generalize the squared binomial rule to the entire product $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$, i.e., all pairs of real numbers a and b . I apologize for that over-detailed presentation of obvious things, but they cannot be so obvious if only a few researchers have noticed that the corresponding procedure might be applied to diagrams as well (see e.g. Foo et al., 1999; Dove, 2002; Giaquinto, 2007).

3.4. FROM PARTICULAR TO GENERAL: THEOREM OF CONSTANTS

Let us therefore try to apply the reasoning with variables shown in the previous section to a diagrammatic proof of the squared binomial formula.

First, let us construct the diagram for the formula, starting by drawing a straight line segment of certain length. Then, instead of putting it against some number axis to measure its particular length, let us assume that the length is unspecified and then take this segment as a *symbol* for segments of *any* length (like the variable a which represents a number of any magnitude).

Let us now colour the segment red, just to differentiate it from another one, which will be coloured blue, be of some other unspecified length and drawn collinear to the red one and positioned so that its beginning coincides with the end of the red one, see Fig. 5(a). We obtain

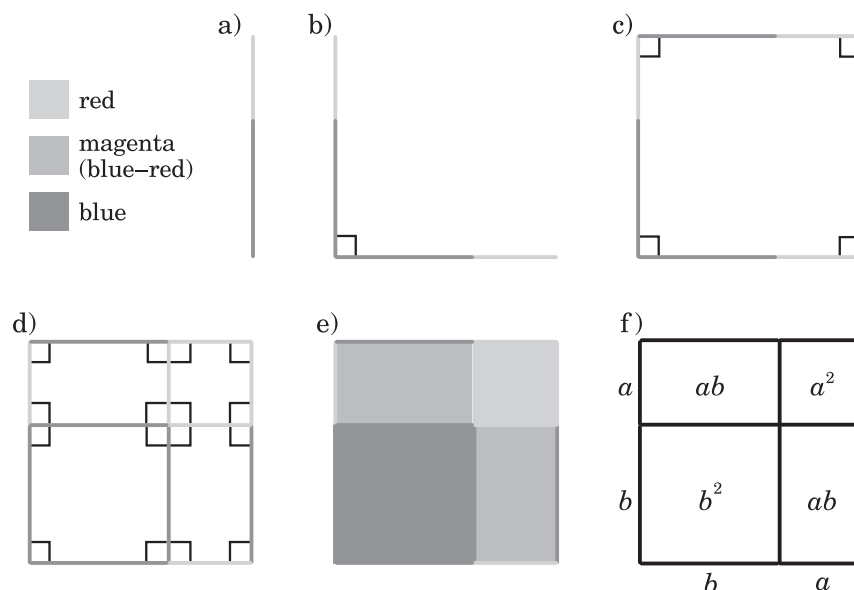


Figure 5. Squared binomial formula proved diagrammatically: two collinear and touching line segments with unspecified lengths (a), copying and rotating the segment by $-\pi/2$ (b)... until the square is built (c), dividing the square (d), and colouring it (e), replacing colours by labels (f). As we could not afford the charge for colour figures imposed by this journal, we were forced to use gray tones in the figure, as explained in the legend.

a longer segment composed of our two segments, aligned and touching. Now use the geometry rule saying that if we compose a segment from two others so arranged, then the length of the composed segment equals the sum of lengths of the constituent segments, *irrespectively of the lengths of the constituent segments*. Now take a copy of our red-blue segment and rotate it around its bottom end by the angle of $-\pi/2$, and then use the rule that the angle between the old segment and its rotated copy will be thus a right angle, as indicated in Fig. 5(b). Now make copies of the two segments and move the copy of the vertical one to the right by its length, and the horizontal one to the top by the same distance. We will obtain the square shown in Fig. 5(c), with the length of the side equal to the sum of lengths of the red and blue segments, and the area equal to the square of the same sum. Then use two more copies of our composite segments and place them at the touching endpoints of the red and blue segments constituting the sides of the square. We will again omit for brevity the geometry rules assuring that the added segments will make right angles with themselves and with the old ones and divide the square into four rectangles with blue and red sides, as shown in Fig. 5(d). We can now colour the insides of the resulting

rectangles with colours corresponding to colours of their sides, getting the figure in Fig. 5(e) (we omitted perpendicularity marks for clarity). The area of this figure is obviously:

The sum of the area of the square with the side length equal to the length of one of the segments, the area of the square with the side length equal to the length of the second of the segments, and areas of the two rectangles with one side length equal to the length of one of the segments and the other side length equal to the length of the other segment.

What does it show us? Because the squares in Fig. 5(c) and 5(e) are equal we obtain the following equality:

The area of a square with lengths of sides that are sums of two given segment lengths *is equal to* [insert here the text from the previous quote].

And remember — all that irrespectively of any particular lengths of the segments involved.

The equality as spelled out above is long and unwieldy. Fortunately, mathematicians long ago invented a much more concise mathematical notation for such occasions (and for others, much worse ones...). For that, let us replace the colours with letter labels, assume that they denote the lengths of the labelled sides and areas of the labelled figures, as in Fig. 5(f). Now we can formulate the equality much more concisely, adding also appropriate quantification, until now assumed implicitly:

$$(\forall a, b \in \mathbb{R})((a + b)^2 = a^2 + 2ab + b^2). \quad (1)$$

Yes, our old squared binomial formula, which we have probably already forgotten during these long and detailed considerations. We could have used the labels instead of colours from the very beginning, but the use of colours is certainly nicer (if one forgets the charge) and more diagrammatic. Recalling that our geometrical reasoning was on every step independent of any concrete lengths of the red and blue segments, we are now assured that the resulting formula is valid for any a and b ,⁵ with no trace of any special generalization problem.

The particularity problem is not restricted to diagrams — the same problem may also occur with formulae. Logicians long ago found a reliable means of generalization from a particular example, used in the

⁵ Of course, we simplified things a little — the geometrical argument above is fully valid for nonnegative numbers only. Extending it to negative numbers requires either some simple algebraic manipulation with the final formula, or another diagrammatic construction. This divergence into two significantly different cases is a common feature of diagrammatic proofs, as signified by the *divergence rule* discussed in Section 3.1.

world of formulae. It is called the Theorem of Constants. It can easily be applied to diagrams too (see e.g. Foo et al., 1999; Giaquinto, 2007). The rule can be in this case formulated as follows:

Theorem of Constants rule:

$$\begin{array}{l} \text{If } (\forall x \in X)Q(x) \\ \text{and } x_0 \in X \ \& \ Q(x_0) \Rightarrow P(x_0), \\ \text{then } (\forall x \in X)P(x), \end{array} \quad (2)$$

provided x_0 does not occur in the formulae $Q(x)$ and $P(x)$, so that the proof of (2) does not use in any way the particular properties of the particular value x_0 .

That is, if all members of the set X have the property Q , and x_0 is a particular object from X , and we can prove that some property P follows from the property Q for this object x_0 in a way which is independent of any other particular properties of the object x_0 , it means that the property P holds for all members of the set X .

In the case of our squared binomial formula, x is a two-dimensional variable $x = (a, b) \in \mathbb{R}^2$. Let us choose $x_0 = (1, 3)$. Then we must prove that $(1 + 3)^2 = 1^2 + 2(1 \cdot 3) + 3^2$ in a way that does not use arithmetic tautologies involving particular properties of 1 and 3, such as $1 + 3 = 4$, $1^2 = 1$, $1 \cdot 3 = 3$, etc. Only properties that are independent of these particular values are allowed in the proof. It means that our proof can use only general tautologies of real arithmetic, the same as those used in the algebraic derivation in Section 3.3, but now applied to the particular $x_0 = (1, 3)$ instead of to an unspecified x . That is:

$$\begin{array}{ll} (1 + 3)^2 = (1 + 3) \cdot (1 + 3) & \text{– by the definition} \\ & \text{of a square power,} \\ = 1 \cdot 1 + 1 \cdot 3 + 3 \cdot 1 + 3 \cdot 3 & \text{– multiplication of sums rule,} \\ = 1^2 + 1 \cdot 3 + 1 \cdot 3 + 3^2 & \text{– squaring and commutativity} \\ & \text{of multiplication,} \\ = 1^2 + 2(1 \cdot 3) + 3^2 & \text{– replacing multiple addition} \\ & \text{by multiplication.} \end{array}$$

Thus:

$$(1 + 3)^2 = 1^2 + 2(1 \cdot 3) + 3^2.$$

Then replacing the particular constant $x_0 = (1, 3)$ by the general variable (a, b) we obtain the formula (1). Now we see why our first “proof” of the formula in Section 3.3, using constants 1 and 3, was wrong and cannot be used as a basis for a universal generalization — we executed there the operations $1 + 3 = 4$, $3^2 = 9$, $1^2 = 1$, etc. which

are true statements of the theory of arithmetic, but they contain the constant $x_0 = (1, 3)$, therefore violating the proviso of the Theorem of Constants, as being dependent on the particular value of x_0 .

In this example, the Theorem of Constants contributes little if anything to our theorem-proving capabilities. This can be different for other cases, say, for diagrams. In diagrams, their particularity means that we do not have at our disposal a diagrammatic equivalent of special letter symbols that may represent any object from some set (that is, no special notation for variables, see Section 3.5). We must use instead graphical objects representing some particular members of the set. The red and blue segments have particular, constant lengths, although we are using them as representing segments of any length. They are constants pretending to be variables — but this is exactly the same as in the proof of the squared binomial formula just conducted above, with the constants 1 and 3 pretending to be variables ranging over all of \mathbb{R} .

Because of that, diagrammatic proofs like that in Fig. 5 are valid, despite using particular lengths of the red (a) and blue (b) segments, provided the diagrammatic reasoning conducted does not rely on the particular lengths of the segments. That is, we must ensure that those elements of the structure of the diagrammatic construction that are used in formulation of the thesis will be the same no matter what particular values the lengths of the segments will take. In this example, it is clear that the diagram will always consist of a larger square divided into two smaller squares and two congruent rectangles with areas as shown. This ensures, according to the Theorem of Constants, that the generalization from this particular diagram to the universally quantified formula (1) is valid.

Now we are finally in a position to pinpoint precisely what kind of error we have made trying to generalize to the set of all triangles the conclusion derived from the diagram in Fig. 3. In the figure we have drawn a particular triangle, which happened to be an acute one. Then the altitudes were drawn, and we proved they intersect in a single point. However, we also decided that this point lies inside the triangle, overlooking the fact that this property is a consequence of the triangle being of the particular type we have drawn (an acute triangle, in which all altitudes lie inside the triangle, therefore so does their intersection). However, in other triangle types some of the altitudes may not lie inside the triangle, so that their intersection may also not lie inside. In summary, we have violated the proviso of the Theorem of Constants, using in the reasoning a particular property of that constant triangle we have actually used in the proof.

3.5. REPRESENTING VARIABLES IN DIAGRAMS

Because a diagram, allegedly, can represent only a *particular* object or configuration of objects, it is not surprising that many researchers in diagrammatics (this author included) repeat the following statement without protest:

“A diagram cannot represent variables, i.e., objects which may denote many different objects (values).”

Again, let us consider how variables are represented in formulae. When we encounter an “ x ” or a “ y ” in a formula, how do we know it is a variable? Not because it is a letter, instead of, say, a numeral. After all, “ π ” is also a letter, but not a variable, only a particular number, just like “ e ” (the base of natural logarithms or Euler’s number $e = 2.7182818284590\dots$) or “ i ” (the imaginary unit $\sqrt{-1}$). However, in other contexts, “ e ” and “ i ” can easily be used as variables. Concerning “ x ” or “ y ,” they are assumed as variables by a conventional interpretation rule (although one rarely explicitly stated): just these letters are commonly used for variables, and it is not recommended to use them for other purposes under the risk of causing confusion.

But what about other letters often used for variables too? Usually, an appropriate context decides. Such contexts include quantification, e.g. $(\exists m)(m \cdot m = 25)$, and the definition of sets, e.g. $E = \{e \mid e = 2n\}$. We can use e here despite its use as Euler’s number because it is a bounded variable within the definition, though it is nevertheless not advisable to use in this vicinity the letter “ e ” as a name of the Euler’s number itself. Also another general interpretation rule is assumed, namely, that all named objects that are not indicated as being constant (e.g., by fixing their values like $x_0 = 1$, $\alpha = \pi/2$ or asserting their constancy by $c = \text{const}$) should be interpreted as variables.

At first sight, it is unclear if that might work for diagrams. Recall that in the conclusions placed at the end of Section 3.4 we asserted that diagrammatic representations *do not* contain a diagrammatic equivalent of special letter symbols that may represent variables directly. The next section shows the ways of overcoming this problem, already commonly exploited in diagrammatic reasoning, although somehow unconsciously, as the often repeated statement in the beginning of this section signifies.

3.6. VARIABLE RECOGNITION RULE

Fortunately, with due respect to those (this author included) repeating the “diagrams cannot represent variables” mantra, it is not true. It does not mean that variables cause no troubles in diagrams, however.

There are generally two main ways a diagram can represent variables. The first one consists of representing them indirectly, according to the general interpretation rule below, while the second way uses hybrid, diagrammatic-propositional representations.

The variable recognition rule. *Any diagrammatic objects or their parameters occurring in a diagram (like points, line segments, figures, positions, lengths, areas, angles, etc.) which are not restricted to be constant by the construction of the diagram or by some explicit graphical statement (including marking them with appropriate symbols or numerical labels), should be interpreted as variable.*

This convention, though rarely stated explicitly, is well known for most of the users of diagrams and assumed by them implicitly. For example, in the squared binomial diagrammatic proof, see Fig. 5(a-e) in Section 3.4, the only candidates for variables are the lengths of red and blue segments. Angles are marked as constant (right angles), while position and orientation of the figure are obviously irrelevant.

The important difference between diagrams and formulae consists here in the fact that the diagrammatic objects considered as variables perform a *dual role* in the diagram. They have the graphical shape of their values (in a sense), so that on the one hand they play the role of symbols of any value, but on the other hand, they may explicitly exhibit the behaviour of their particular values. This feature, although it greatly assists in visual understanding of semantics of the problem and creating proper intuitions about it, carries with it the danger of assuming that all other elements of the set spanned by the variable behave in the same way as that particular value shown (i.e., lead to the same diagram structure and reasoning process). In this way we may use some hidden, additional assumptions in the reasoning which are valid for these particular objects, but not necessarily shared by all members of the assumed set, thus violating the proviso of the Theorem of Constants in Section 3.4 (see also the final analysis of the Altitudes in a Triangle theorem there).

Besides this strictly diagrammatic, though indirect means, variables in diagrams can be represented explicitly as in formulae, with symbolic letters in textual labels and formulae embedded in hybrid diagrammatic-propositional representations commonly used, as shown in the squared binomial diagram in Fig. 5(f). Of course, as they label particular diagram elements, this may also lead to the particularity error if we are careless.

Moreover, knowing that something is a variable does not suffice. To be able to use the object sensibly in reasoning, we must also specify what values this variable can assume, i.e., we should specify the set

over which it varies. The ways of doing this in diagrams are presented in Section 3.8.

3.7. REPRESENTATION OF QUANTIFIERS IN DIAGRAMS

Superficially, it also seems true that we cannot represent quantifiers in a diagram. Indeed:

There are no common diagrammatic symbols or constructions for quantifiers, and the two-dimensional structure of diagrams prevents easy delineation of the quantified statement and its association with the appropriate variable.

However, again there are diagrammatic interpretation rules providing functionality analogous to quantification.

First, let us observe that the main element of quantification is, as for the specification of variables (see Section 3.5), the representation of a set over which the quantification is made.

What is additionally needed, is the way of distinguishing between universal and existential quantifiers, which is usually done by an interpretation convention. When we demonstrate that the reasoning depicted in a diagram is valid for all members of some set (see Section 3.8), we have a universal quantification. Otherwise, when we only show the construction for some particular object, we only demonstrate the existence of an object with the appropriate property. Due to diagram imprecision, there remains a problem of ensuring that we have not got the case of impossible configuration, see Section 2.1. Other ways of distinguishing such implicit quantifiers were proposed by Winterstein (2004).

In summary, quantification in diagrams is in most cases also an artificial problem, the real problem being again the representation of sets instead. There are many ways of representing sets in diagrams, see Section 3.8 for details.

3.8. REPRESENTING SETS IN DIAGRAMS

When discussing the problems of generalization, particularity, variables representation, and quantifier representation, we could observe that the various ways of solving these problems eventually boil down to the problem of representation of some sets in a diagram. There are many ways of representing sets in diagrams, some of them were already briefly introduced. One may generally distinguish five such ways:

- *variable* recognition and *interpretation rules* for implicit specification of the sets the variables are varying over;
- definition with *formulae in hybrid representations*;

- explicit representation of set elements or subsets *by divergence* into cases;
- dynamic diagrams showing a recipe for implicit *generation of the set*;
- explicit representation of sets (as *graphical objects in a diagram*).

They are often mixed together in various ways in actual diagrams.

3.8.1. *Variable interpretation*

With the variable recognition rule of Section 3.6 comes also a set of rules indicating the proper set over which the given variable value may range, providing thus an indirect representation of quantification as well. For example, lengths of sides of the figures and their areas are assumed to range over positive reals (often allowing also for the “degenerate case” of 0). Angles, depending on the context, are assumed to range over the intervals $[0, \pi]$ or $[0, 2\pi]$, etc. Although these rules remain for the most part uncodified, diagram users know them well, as attested e.g. by (Winterstein et al., 2000, Section 5.1). The authors formulated the rules as allowable transformations that do not invalidate the given diagrammatic argument and an informal survey among students has shown that they have had little trouble in choosing the proper transformations. What remains to be done is the systematic codification of these rules, in order to develop a standard to be used in normal circumstances, together with allowable ways of informing the user when some deviation from the standard is needed in a given diagram.

3.8.2. *Hybrid representation*

The commonly used hybrid representations contain a diagram which usually represents the structure of the problem, while textual labels, associated with specific graphical components reinforce the variable interpretation and fix the correspondence between the diagrammatic representation and the usually propositional formulation of premises and conclusions of the reasoning. The drawing can also incorporate formulae, providing for example precise specification of sets spanned by diagrammatic variables or relations between elements, especially when standard interpretation rules must be modified. With formulae we can explicitly write down the specification of the required set. An example is the definition of the set $E = \{e \mid e = 2n\}$ which appeared already as an example, though it should be augmented by the specification of the variable n , say $n \in \mathbb{N}$, or more explicitly, $n \in \{0, 1, 2, \dots\}$. At other times implicit specifications are assumed, according to the context. E.g., in texts on real analysis variables like x or y are assumed to always vary over the set of real numbers \mathbb{R} , without a need for explicit specification

$x, y \in \mathbb{R}$, while variables like i, j, k or n are usually assumed to vary over the set of natural numbers \mathbb{N} (or integers \mathbb{Z}), etc.

An example of using these ways in diagrams is provided by the arithmetic theorem ascribed to Nicomachus of Gerasa, see (Nelsen, 1993). Figure 6(a) is a simplified version of the diagram from that book.

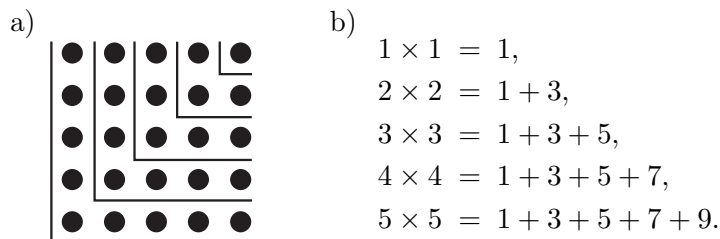


Figure 6. The Nicomachus diagram, adapted from (Nelsen, 1993, p. 71), in simplified form (a), and the particular equalities represented (b).

This diagram is a nice example of the “particularity problem” discussed in Section 3.3, as in this form it represents only a few particular equalities, see Fig. 6(b).

At first sight it seems hard to find a variable or variables here. But taking into account the represented equalities it becomes obvious that the variable is actually the whole diagram, parameterized by its size (say, side length), let us call it n , varying over \mathbb{N} (the set of naturals). Generalizing the argument to any n must be, however, done outside the diagram, with appropriate reasoning starting from the particular formulae extracted from the diagram. We can, however, more explicitly show the variable of the problem and a way to generalize the reasoning to arbitrary $n \in \mathbb{N}$ by drawing a more elaborate diagram, as in Fig. 7(a). Finally, we may produce a hybrid representation of the problem in Fig. 7(b). In this diagram textual labels clarify the reasoning even further and fix the correspondence between the diagrammatic proof and the propositional formulation of the theorem.

3.8.3. Divergence

When the set is finite (hence discrete), it is possible to prove the theorem by repeating the reasoning, with necessary modifications, for every member of the set — i.e., by using divergence. It is not so rare as one might think: see for example reasoning conducted using logical tables, like the Karnaugh map. A discrete and infinite example is given by the Nicomachus diagram. Here also the reasoning is repeated for some finite subset of the set (which here consists of squares of dots with different sizes). From these cases the pattern of change between the cases is extracted (here, it is the formula $2n - 1$ for the general component

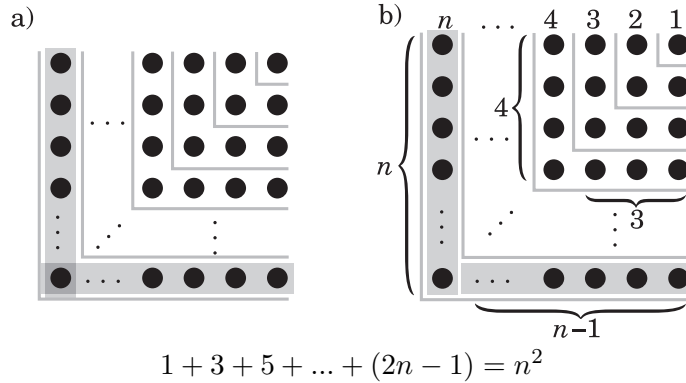


Figure 7. The Nicomachus diagram with diagrammatic variables made visible (a), and its hybrid version with propositional constants and variables added as labels (b).

of the sum). For continuous domains the set is often infinite, but not discrete. Here a finite divergence is obtained, with cases corresponding to (infinite) subsets of the set, including single, border cases (the degenerate instances), if any. The situation may be illustrated by the Altitudes in a Triangle problem of Section 3.2. Here we have three cases for three infinite subsets of the set of all triangles. For completeness, one may add the case of the equilateral triangle, where the intersection point has some additional properties (as coinciding with all other centerpoints of the triangle). This fourth subset contains only a single element, if we factor the set by shape-preserving transformations.

3.8.4. *Dynamic diagram*

Yet another method consists of the use of so-called “dynamic diagrams.” This does not mean animation, although they can be animated too. The idea is instead to show not only a diagrammatic construction for some particular case, but also a kind of recipe for generating diagrams for other members of the set.

A simple example was given by (Arnheim, 1969), see Fig. 8. A particular triangle is drawn, with an easy diagrammatic proof that all its internal angles sum up to 180° . At the same time, graphical indicators at the ends of extended sides of the triangle (together with the parallelism indicator) supply the necessary structure variation argument showing that whatever the directions of the sides may be (that is, effectively for all triangles) the relationship proven for this particular triangle will remain unchanged.

This technique can be made interactive (King and Schattschneider, 1997; Lindsay, 2000; Le and Kulpa, 2003; Le and Kulpa, 2004; Otte,

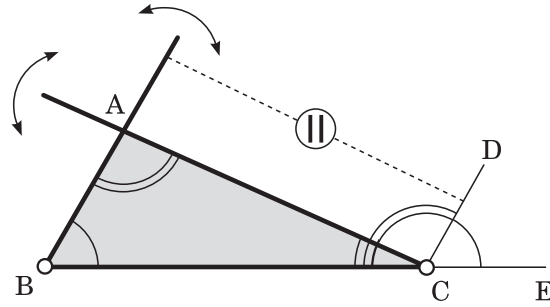


Figure 8. A diagrammatic proof for the sum of angles in a triangle, using a dynamic diagram (adapted from (Arnheim, 1969), extended).

2006, p. 147) which greatly facilitates the process of finding the proper set over which our “ x ” ranges.

3.8.5. *Explicit representation of sets*

In other cases we may explicitly represent the set involved, if it can be represented as some line or figure in the plane. One of the most common cases involves the use of Cartesian coordinate axes, which are in fact number axes representing the set of real numbers \mathbb{R} , or possibly the set of integers \mathbb{Z} . Graphs of functions constitute in this case one dimensional sets of pairs of numbers. How that works, can be shown with a simple theorem of reciprocal inequality.

If a real number x is positive, then $x + 1/x \geq 2$. The premises of this statement can be written as:

$$x \in \mathbb{R} \ \& \ x > 0.$$

After adding the necessary quantification for completeness, the statement can be formulated with a single formula:

$$(\forall x)(x \in \mathbb{R} \ \& \ x > 0) \Rightarrow x + 1/x \geq 2. \quad (3)$$

The particular set occurring in this example can also be explicitly referred to by giving it a name:

$$X = \{x \mid x \in \mathbb{R} \ \& \ x > 0\}, \quad (4a)$$

$$(\forall x \in X) x + 1/x \geq 2. \quad (4b)$$

Now let us see how we can use the direct representation of the set X . The standard diagrammatic representation of the set of positive reals X is the positive half of the real number axis, see Fig. 9. An empty circle at $x = 0$ denotes that this value does not belong to X .

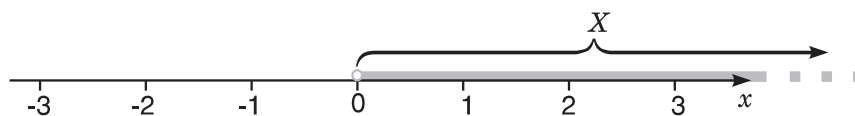


Figure 9. A number axis.

Then, we must represent our inequality to be proved. In the left hand side it contains a sum of two functions: x and $1/x$. The graphs representing these functions are shown in Fig. 10(a), with three characteristic points for $x = 1/2, 1$, and 2 shown explicitly.

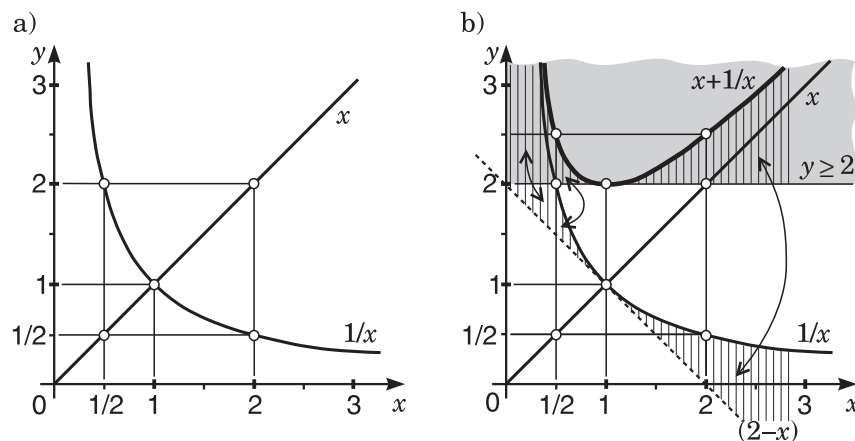


Figure 10. Proving the reciprocal inequality with explicit representation of the set X : graphs of functions x and $1/x$ (a); proving the inequality (b).

Representations of the sum of these two functions, as well as the right hand side of the inequality, are shown in Fig. 10(b). The theorem will be proven when we demonstrate that the graph of the sum (the heavy curve) must lie entirely in the gray area. This can be done as follows. The sum $x + 1/x$ equals 2 at $x = 1$, and the slope of the tangent to the graph of $1/x$ equals -1 at this point. For $x > 1$, the function $1/x$ decreases, but the function x increases faster than $1/x$ (because the tangent to $1/x$ has its slope always smaller than for $x = 1$). Therefore, the sum of the functions will increase too, staying thus always larger than the value 2 attained at $x = 1$. Symmetrical situation occurs for $x < 1$ — here x decreases, but $1/x$ increases faster, hence the sum will also increase, keeping the value of the sum larger than 2. So the theorem is proven.

Another way of looking at the diagram starts from the dotted line (with the equation $2 - x$). Its sum with the function x is a constant 2, i.e., the line $y = 2$. Because the graph of $1/x$ lies entirely above the

dotted line, the sum of $x + 1/x$ will always lie above the line $y = 2$, as indicated additionally by thin vertical lines.

Yet another version of this proof, also using a function graph diagram, was provided by Nelsen (1993).

Marking the variable x as a label of the number axis indicates that it ranges over the set of (positive) reals explicitly represented as (a positive half of) the axis (strictly speaking a finite fragment of the infinite half-line). The diagram thus shows that the required relations between the functions hold effectively *for all* elements of X , thus representing implicitly also the universal quantification. The variable appears then within formulae summarizing propositionally the conclusion from the diagram.

The set X is here not only infinite, but also unbounded, hence it cannot be represented in a bounded diagram in its entirety. In such cases the missing part is often explicitly indicated by an ellipsis symbol “...” (see the large gray dots on the right in Fig. 9). For coordinate axes (indicated by arrows at the positive ends of the axes), the ellipses at both ends are usually omitted by convention; it is assumed implicitly that the axes extend to infinity at both ends. Such “abstraction devices,” as they are called (Jamnik, 2001), are commonly used in reasoning with formulae (ellipsis, summation sign \sum , etc.), but are seldom used in rigorous diagrammatic reasoning as yet, and are even explicitly disallowed by some authors (Jamnik, 2001). This is due to the lack of precise semantics and rules of proper usage for them in diagrammatic applications, especially in automatic reasoning systems. They can thus lead to errors and ambiguities (Jamnik, 2001). However, an elegant and rigorous semantics of one kind of a single ellipsis has already been given in (Foo et al., 1999). As there are no significant changes of behaviour of the functions x and $1/x$ for large x 's, here the ellipsis does not lead to problems. The same remark is valid for the y axis. The ellipsis can also be troublesome in formulae, and requires an “intelligent” parser to interpret it rigorously in reasoning involving lists (Bundy and Richardson, 1999). Another work addressing this problem, for matrix block notation, which is closer to diagrams than list notation, appeared in (Pollet et al., 2004).

In other cases, various sets can be represented by two-dimensional figures, provided they can be parameterized by pairs of real numbers. For example, several sets of *all triangles* (triangle spaces) have been developed in (Kulpa, 2009b, forthcoming). The use of such spaces greatly facilitates finding the proper generalization by the divergence method for problems like Altitudes in a Triangle in Section 3.2.

4. Conclusions

Diagrammatic reasoning differs significantly from reasoning with formulae, so that it requires the development of new techniques for conducting reliable and rigorous reasoning. For some problems from the problem cluster concerned with universal generalization in diagrams, it nevertheless became possible to transfer to diagrams a few such techniques from ordinary mathematical reasoning, after appropriate adaptation, like the divergence method (Section 3.2), and the Theorem of Constants (Section 3.4).

However, as expected, some specifically diagrammatic features appeared during the search for solutions of problems of this cluster. They include the specific variable recognition rules, commonly used by reasoners using diagrams, but nowhere satisfactorily codified as yet (Section 3.6), similar rules for finding sets over which the variables vary and quantifiers range (Section 3.8), and the peculiar property of diagrammatic variables, which perform a dual role of both variables and constants. The latter property is, on the one hand, very useful for better understanding of the semantics of the reasoning and building proper intuitions for searching the correct route to the proof, but, on the other hand, it increases the danger of conducting a particularity error, by facilitating the violation of the proviso of the Theorem of Constants (Section 3.6).

Diagrammatic techniques of representing sets in diagrams are also very useful for solving the generalization problem. We have here several indirect and direct methods, discussed in some detail in Section 3.8. One of the methods, namely representing sets directly as lines and figures in a diagram, with several distinct constructions for the set of all triangles, will be discussed in more detail in a separate paper (Kulpa, 2009b, forthcoming).

Another important cluster of diagrammatic problems, namely the “diagram imprecision” problem, has been briefly presented in the paper as well. It is harder to resolve than the generalization problem, as it has practically no analogue in the world of formulae. Its more thorough discussion will be presented in a separate paper (Kulpa, 2009a, forthcoming).

Only after conducting such detailed analysis of all the problems interfering with attempts at rigorous reasoning using diagrams, and after finding ways of avoiding various traps and errors that await us there, will we be able to put diagrammatic reasoning on secure foundations and use it as a reliable and rigorous tool of mathematical reasoning, on a par with already well developed techniques using formulae. We hope this paper will prove to be a useful contribution to this enterprise.

Author's Vitae

Kulpa, Z.

Dr. Kulpa holds the position of Associate Professor in the Institute of Fundamental Technological Research of the Polish Academy of Sciences (IPPT PAN) in Warsaw, Poland. He graduated from Department of Electronics of the Warsaw Technical University, and started his research work in the Institute of Automatic Control of the Polish Academy of Sciences in Warsaw. He obtained his Ph.D. degree from the Institute of Computer Science of the Polish Academy of Sciences in Warsaw, and the D.Sc. degree from the Institute of Fundamental Technological Research. His research interests started from computer synthesis of switching circuits, then moved to computer image processing and recognition, computer graphics (especially graphical man-machine interfaces), and finally qualitative physics. Here he became interested in interval computations and developed a comprehensive system of diagrammatic representation and reasoning for interval analysis (see his recent book at <http://www.ippt.gov.pl/~zkulpa/diagrams/diawa.html>). This has led him to his current research interest, namely diagrammatics (or diagrammatic representation and reasoning).

Acknowledgements

The research leading to this paper was partially supported by the Research Project No. 5 T07F 002 25 (for the years 2003–2006), granted by KBN (State Committee for Scientific Research).

References

- Arnheim, R.: 1969, *Visual Thinking*. University of California Press, Berkeley, CA.
- Bundy, A., Richardson, J.: 1999, Proofs about lists using ellipsis. In: Ganzinger, H., McAllester, D., Voronkov, A., eds.: 1999, *Logic for Programming and Automated Reasoning. LNAI 1705*, Springer-Verlag, Berlin, pp. 1–12.
- DIAGRAMS: 1992, *Reasoning with Diagrammatic Representations (1992 AAAI Spring Symposium)*. AAAI Press, Menlo Park, CA.
- Dove, I.: 2002, Can pictures prove? *Logique & Analyse*, 179-180: pp. 309–340.
- Dubnov, Y.S.: 1955, *Oshibky v geometricheskykh dokazatelstvakh* (in Russian). GITTL, Moscow. [English translation: Dubnov, Y.S.: 1963, *Mistakes in Geometric Proofs*. Heath, Boston, MA.]
- Foo, N.Y., Pagnucco, M., Nayak, A.C.: 1999, Diagrammatic proofs. In: *Proc. 16th International Joint Conference on Artificial Intelligence (IJCAI-99)*, Stockholm, Aug 1999. Morgan Kaufmann, pp. 378–383.

- Frege, G.: 1879, *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. Halle a. S. [English edition: Frege, G.: 1879, *Begriffsschrift, A Formula Language, Modeled Upon That of Arithmetic, For Pure Thought*. Reprinted in: Jean van Heijenoort, ed.: 1967, *From Frege to Gödel: A Source Book in Mathematical Logic*. Harvard University Press, Cambridge, MA, pp. 1879-1931.]
- Giaquinto, M.: 2007, *Visual Thinking in Mathematics: An Epistemological Study*. Oxford University Press, Oxford.
- Greaves, M.: 2001, *The Philosophical Status of Diagrams*. CSLI Publications, Stanford, CA.
- Jamnik, M.: 2001, *Mathematical Reasoning with Diagrams: From Intuition to Automation*. CSLI Publications, Stanford, CA.
- King, J.R., Schattschneider, D., eds.: 1997, *Geometry Turned On!: Dynamic Software in Learning, Teaching, and Research*. Mathematical Association of America, Washington, DC.
- Kulpa, Z.: 1983, Are impossible figures possible? *Signal Processing*, 5(3): pp. 201–220.
- Kulpa, Z.: 1987, Putting order in the impossible. *Perception*, 16: pp. 201–214.
- Kulpa, Z.: 2003a, Self-consistency, imprecision, and impossible cases in diagrammatic representations. *Machine GRAPHICS & VISION*, 12(1): pp. 147–160.
- Kulpa, Z.: 2003b, *From Picture Processing to Interval Diagrams*. IFTR Reports 4/2003, Warsaw, 313 pp.
[See <http://www.ippt.gov.pl/~zkulpa/diagrams/fpptid.html>]
- Kulpa, Z.: 2009a, Main problems of diagrammatic reasoning: Part II: The imprecision problem, (in preparation).
- Kulpa, Z.: 2009b, Representing sets in diagrams: Spaces of all triangles, (in preparation).
- Le, T.L., Kulpa, Z.: 2003, Diagrammatic spreadsheet. *Machine GRAPHICS & VISION*, 12(1): pp. 133–146.
- Le, T.L., Kulpa, Z.: 2004, Diagrammatic spreadsheet: An overview. In: Blackwell, A., Marriott, K., Shimojima, A., eds.: 2004, *Diagrammatic Representation and Inference*. LNAI 2980, Springer-Verlag, Berlin, pp. 420–423.
- Lemon, O., Pratt, I.: 1997, Spatial logic and the complexity of diagrammatic reasoning. *Machine GRAPHICS & VISION*, 6(1): 77–88.
- Lindsay, R.K.: 2000, Playing with diagrams. In: Anderson, M., Cheng, P., Haarslev, V., eds.: 2000, *Theory and Application of Diagrams*. LNAI 1889. Springer Verlag, Berlin, pp. 300–313.
- Luengo, I.: 1995, *Diagrams in Geometry*. Ph.D. Thesis, Indiana University, Bloomington, IN.
- MacInnis, R., McKinna J., Parsons, J., Dyckhoff. R.: A mechanised environment for Frege's *Begriffsschrift* notation. Proc. Mathematical User-Interfaces Workshop 2004, Bialowieza, Poland, Sept. 18, 2004.
[See <http://www.activemath.org/~paul/MathUI04/proceedings/>]
- Maxwell, E.A.: 1959, *Fallacies in Mathematics*. Cambridge University Press, Cambridge.
- Miller, N. : 2001, *A Diagrammatic Formal System for Euclidean Geometry*. Ph.D. Thesis, Cornell University, Ithaca, NY.
- Muzalewski, M.: 1993, *An Outline of PC Mizar*. Foundation Philippe le Hodey, Brussels. [See <http://mizar.org/project/bibliography.html>]
- Needham, T.: 1997, *Visual Complex Analysis*. Clarendon Press, Oxford.

- Nelsen, R.B.: 1993, *Proofs Without Words: Exercises in Visual Thinking*. The Mathematical Association of America, Washington, DC.
- Otte, M.: 2006, Proof-analysis and continuity. *Foundations of Science*, 11: pp. 121–155.
- Pollet, M., Sorge, V., Kerber, M.: 2004, Intuitive and formal representations: The case of matrices. In: Asperti, A., Bancerek, G., Trybulec, A., eds.: 2004, *Mathematical Knowledge Management. LNCS 3119*, Springer-Verlag, Berlin, pp. 317–331.
- Shin, S.-J.: 1994, *The Logical Status of Diagrams*. Cambridge University Press, Cambridge, MA.
- Winterstein, D., Bundy, A., Jamnik, M.: 2000, A proposal for automating diagrammatic reasoning in continuous domains. In: Anderson, M., Cheng, P., Haarslev, V., eds.: 2000, *Theory and Application of Diagrams. LNAI 1889*, Springer-Verlag, Berlin, 286–299.
- Winterstein, D.: 2004, *Using Diagrammatic Reasoning for Theorem Proving in a Continuous Domain*. Ph.D. Thesis, The University of Edinburgh, Edinburgh, 263 pp.