Algorithms and Geometric Constructions

Vladimir Uspenskiy[1] and Alexander Shen[2(✉)]

[1] Moscow State Lomonosov University, Moscow, Russia
[2] LIRMM CNRS and University of Montpellier, on leave from IITP RAS, Montpellier, France
alexander.shen@lirmm.fr

Abstract. It is well known that several classical geometry problems (e.g., angle trisection) are unsolvable by compass and straightedge constructions. But what kind of object is proven to be non-existing by usual arguments? These arguments refer to an intuitive idea of a geometric construction as a special kind of an "algorithm" using restricted means (straightedge and/or compass). However, the formalization is not obvious, and different descriptions existing in the literature are far from being complete and clear. We discuss the history of this notion and a possible definition in terms of a simple game.

1 Introduction

The notion of an algorithm as an intuitively clear notion that precedes any formalization, has a rather short history. The first examples of what we now call algorithms were given already by Euclid and al-Khwârizmî. But the general idea of an algorithm seems to appear only in 1912 when Borel considered "les calculus qui peuvent être réellement effectués"[1] and emphasized: "Je laisse intentionnellement de côté le plus ou moins grande longeur pratique des opérations; l'essentiel est que chaque de ces opérations soit exécutable en un temps fini, par une méthode sûre et sans ambiguïte"[2] [4, p. 162]. The formal definition of a representative class of algorithms was given in 1930s (in the classical works of Gödel, Church, Kleene, Turing, Post and others); the famous Church–Turing thesis claims that the class of algorithms provided by these formal definitions is representative.

In this paper we look at the history of another related notion: the notion of a *geometric construction*. One may consider geometric constructions as a special type of algorithms that deal with geometric objects. Euclid provided many examples of geometric constructions by compass and straightedge (ruler); later these constructions became a standard topic for high school geometry exercises. Several classical problems (angle trisection, doubling the square, squaring the

A. Shen—Supported by ANR-15-CE40-0016-01 RaCAF grant.

[1] The computations that can be really performed.

[2] I intentionally put aside the question of bigger or smaller practical length of the operation; it is important only that each of the operations can be performed in a finite time by a clear and unambiguous method".

F. Manea et al. (Eds.): CiE 2018, LNCS 10936, pp. 410–420, 2018.
https://doi.org/10.1007/978-3-319-94418-0_41

circle) were posed and remained unsolved since ancient times (though solutions that involve more advanced instruments than compass and straightedge were suggested). These problems were proved to be unsolvable in 19th century. One would expect that the proof of unsolvability assumes as a prerequisite a rigorously defined notion of a "solution" that does not exist. Recall that the first undecidability proofs could appear only after an exact definition of an algorithm was given.

However, historically this was not the case and the impossibility proofs appeared without an exact definition of a "geometric construction". These proofs used the algebraic approach: For example, to show that the cube cannot be doubled, one proves that $\sqrt[3]{2}$ cannot be obtained from rationals by arithmetic operations and square roots. The reduction from a geometric question to an algebraic one looks quite obvious and was omitted by Wantzel who first proved the impossibility of angle trisection and cube doubling. As he wrote in [22], "pour reconnaitre si la construction d'un problème de Géometrie peut s'effectuer avec la règle et le compas, if faut chercher s'il est possible de faire dépendre les racines de l'equation à laquelle il conduit de celles d'un système d'équations du second degré".[3] This is said in the first paragraph of the paper and then he considers only the algebraic question.

Several other interesting results were obtained in 19th century. It was shown that all constructions by compass and straightedge can be performed using the compass only (the *Mohr–Mascheroni theorem*) if we agree that a line is represented by a pair of points on this line. Another famous result from 19th century, the *Poncelet–Steiner theorem*, says that if a circle with its center is given, then the use of compass can be avoided, straightedge is enough. Other sets of tools were also considered, see, e.g., [3,9,14].

Later geometric construction became a popular topic of recreational mathematics (see, e.g., [6,10,13,15]). In most of the expositions the general notion of a geometric construction is still taken as granted, without a formal definition, even in the nonexistence proofs (e.g., when explaining Hilbert's proof that the center of a circle cannot be found using only a straightedge [6,10,15]; see below Sect. 6 about problems with this argument). Sometimes a definition for some restricted class of geometric construction is given (see, e.g., [18]). In [13] an attempt to provide a formal definition is made, still it remains ambiguous with respect to the use of "arbitrary points" (see Sect. 4). Baston and Bostock [2] observe that the intuitive idea of a "geometric construction" has no adequate formal definition and discuss several examples but do not attempt to give a formal definition that is close to the intuitive notion. It seems that even today people still consider the intuitive notion of a "geometric construction algorithm" as clear enough to be used without a formal definition (cf. [1], especially the first `arxiv` version).

In Sect. 2 we consider a naïve approach that identifies constructible points with the so-called "derivable" points. Then in Sects. 3 and 4 we explain why

[3] To find out whether a geometric problem can be solved by straightedge and compass construction, one should find whether it is possible to reduce the task of finding the roots of the corresponding equation to a system of equations of second degree.

this approach contradicts our intuition. In Sect. 5 we suggest a more suitable definition, and finally in Sect. 6 we note that the absence of formal definitions has led to incorrect proofs.

2 Derivable Points and Straight-Line Programs

At first it seems that the definition of a geometric construction is straightforward. We have three classes of geometric objects: points, lines and circles. Then we consider some operations that can be performed on these objects. We need to obtain some object (the goal of our construction) applying the allowed operations to given objects. As Tao [18] puts it,

> Formally, one can set up the problem as follows. Define a configuration to be a finite collection $\mathcal{C}$ of points, lines, and circles in the Euclidean plane. Define a construction step to be one of the following operations to enlarge the collection $\mathcal{C}$:
>
> – (Straightedge) Given two distinct points A, B in $\mathcal{C}$, form the line $\overline{AB}$ that connects A and B, and add it to $\mathcal{C}$.
> – (Compass) Given two distinct points A, B in $\mathcal{C}$, and given a third point O in $\mathcal{C}$ (which may or may not equal A or B), form the circle with centre O and radius equal to the length $|AB|$ of the line segment joining A and B, and add it to $\mathcal{C}$.
> – (Intersection) Given two distinct curves γ, γ' in $\mathcal{C}$ (thus γ is either a line or a circle in $\mathcal{C}$, and similarly for γ'), select a point P that is common to both γ and γ' (there are at most two such points), and add it to $\mathcal{C}$.
>
> We say that a point, line, or circle is constructible by straightedge and compass from a configuration $\mathcal{C}$ if it can be obtained from $\mathcal{C}$ after applying a finite number of construction steps.

We can even try to define the geometric construction algorithm as a straight-line program, a sequence of assignments whose left-hand side is a fresh variable and the right-hand side contains the name of the allowed operation and the names of objects to which this operation is applied.

Baston and Bostock [2] use the name "derivable" for objects that can be obtained in this way starting from given objects. In other words, starting with some set of given objects, they consider its closure, i.e., the minimal set of objects that contains the given ones and is closed under allowed operations. The objects that belong to this closure are called *derivable* from the given ones. In these terms, the impossibility of trisecting the angle with the compass and the straightedge can be stated as follows: *for some points A, B, C the trisectors of the angle BAC are not derivable from* $\{A, B, C\}$.

Baston and Bostock note that the intuitive notion of a "constructible" point (that they intentionally leave without any definition) may differ from the formal notion of a derivable point in both directions. We discuss the differences in the following sections.

3 Uniformity and Tests

There are some problems with this approach. First of all, this appoach is "non-uniform". Asking a high school student to construct, say, a center of an inscribed circle of a triangle ABC, we expect the solution to be some specific construction that works *for all triangles*, not just the proof that this center is always derivable from A, B, and C. The naïve approach would be to ask for a straight-line program that computes this center starting from A, B, and C. However, an obvious problem arises: the operation of choosing an intersection point of two curves is non-deterministic (we need to choose one of two intersection points). We may guarantee only that *some* run of the program produces the required object, or guarantee that the required object is among the objects computed by this program. This is a common situation for classical constructions. For example, the standard construction of the centre of the incircle of a triangle can also produce centres of excircles (the circles outside the triangle that touch one of its sides and the extensions of two other sides).

The non-deterministic nature of the operations was mentioned by different authors. Bieberbach [3] says that the constructions should be performed in the "oriented plane" (not giving any definitions). Tietze [19–21] notes that some objects can be constructed but only in a non-deterministic way, again without giving definition of these notions.

One could give up and consider the non-uniform setting only. As Manin [13, p. 209] puts it, "we ignore how to choose the required point from the set of points obtained by the construction". Another approach is to replace straight-line programs by decision trees where tests appear as internal nodes. Still none of these two approaches (decision trees or non-deterministic choice) is enough to save some classical constructions in a uniform setting as observed by Baston and Bostock [2, p. 1020]. They noted that the construction from Mohr–Mascheroni theorem allows us to construct the intersection point of two intersecting lines AB and CD (given A, B, C, D) using only a compass. Each use of the compass increases the diameter of the current configuration at most by an $O(1)$-factor, and the intersection point can be arbitrarily far even if A, B, C, D are close to each other, so there could be no *a priori* bound on the number of steps. The necessity of an iterative process in the Mohr–Mascheroni theorem was earlier mentioned in another form by Dono Kijne [11, ch. VIII, p. 99]; he noted that this result depends on Archimedes' axiom.

To save the Mohr–Mascheroni construction, one may consider programs that allow loops. This was suggested, e.g., by Engeler [7]. Here we should specify what kind of data structures are allowed (e.g., whether we allow dynamic arrays of geometric objects or not). In this way we encounter another problem, at least if we consider straightedge-only constructions on the *rational* plane $\mathbb{Q}^2$ and allow using tests and do not bound the number of steps/objects. Baston and Bostock [2] observed that having four different points $A, B, C, D \in \mathbb{Q}^2$ in a general position (no three points lie on a line, no two connecting lines are parallel), we can enumerate all (rational) points and therefore all rational lines. Then we can wait until a line parallel to AB appears (we assume that we may

test whether two given lines intersect or are parallel) and then use this parallel line to find the midpoint of AB. This construction does not look like a intuitively valid geometric construction and contradicts the belief that one cannot construct the midpoint using only a straightedge, see [2] for details.

4 Arbitrary Points

Let us now consider the other (and probably more serious) reason why the notion of a derivable object differs from the intuitive notion of a constructible object. Recall the statement about angle trisection as stated by Tao [18]: for some triangle ABC the trisectors of angle BAC are not derivable from $\{A, B, C\}$. (Tao uses the word "constructible", but we keep this name for the intuitive notion, following [2].) Tao interprets this statement as the impossibility of angle trisection with a compass and straightedge, and for a good reason.

On the other hand, the center of a circle is not derivable from the circle itself, for the obvious reason that no operation can be applied to enlarge the collection that consists only of the circle. Should we then say that the center of a given circle cannot be constructed by straightedge and compass? Probably not, since such a construction is well known from the high school. A similar situation happens with the construction of a bisector of a given angle (a configuration consisting of two lines and their intersection point).

Looking at the corresponding standard constructions, we notice that they involve another type of steps, "choosing an arbitrary point" (on the circle or elsewhere). But we cannot just add the operation "add an arbitrary point" to the list of allowed operations, since all points would become derivable. So what are the "arbitrary points" that we are allowed to add? Bieberbach [3, p. 21] speaks about "Punkte, über die keine Angaben affiner oder metrischer Art gemacht sind"[4] and calls them "willkürliche Punkte"—but this hardly can be considered as a formal definition.

Tietze [21] notes only that "the role of arbitrary elements is not so simple as it is sometimes thought". Baston and Bostock [2] explain the role of arbitrary elements, but say only that "the distinction between constructibility and derivability arising from the use of arbitrary points is not very complex" and "we will not pursue a more detailed analysis in this direction"; they refer to [12] for an "elementary approach", but this book also does not give any clear definition. Probably the most detailed explanation of the role of arbitrary points is provided by Manin [13], but he still defines the construction as a "finite sequence of steps" (including the "arbitrary choices") and says that a point is constructible if there exists a construction that includes this point "for all possible intermediate arbitrary choices"; this definition, if understood literally, makes no sense since different choices leads to different constructions. Schreiber [16] tries to define the use of arbitrary points in a logical framework, but his exposition is also far from being clear.

[4] Points for which we do not have affine or metric information.

How can we modify the definitions to make them rigorous? One of the possibilities is to consider the construction as a strategy in some game with explicitly defined rules. We discuss this approach in the next section.

5 Game Definition

The natural interpretation of the "arbitrary choice" is that the choice is made by an adversary. In other words, we consider a game with two players, Alice and Bob. We start with the non-uniform version of this game.

Let E be some finite set of geometric objects (points, lines, and circles). To define which objects x are constructible starting from E, consider the following full information game. The *position* of the game is a finite set of geometric objects. The initial position is E. During the game, Alice and Bob alternate. Alice makes some requests, and Bob fulfills these requests by adding some elements to the current position. Alice wins the game when x becomes an element of the current position. The number of moves is unbounded, so it is possible that the game is infinite (if x never appears in the current position); in this case Alice does not win the game.

Here are possible request types.

- Alice may ask Bob to add to the current position some straight line that goes through two different points from the current position.
- Alice may ask Bob to add to the current position a circle with center A and radius BC, if A, B, C are points from the current position.
- Alice may ask Bob to add to the current position one or two points that form the intersection of two different objects (lines or circles) that already belong to the current position.

If we stop here, we get exactly the notion of derivable points, though in a strange form of a "game" where Bob has no choice. To take the "arbitrary" points into account, we add one more operation:

- Alice specifies an open subset of the plane (say, an open circle), and Bob adds some point of this subset to the current position.

The point x is *constructible* from E if Alice has a winning strategy in this game.

Let us comment on the last operation.

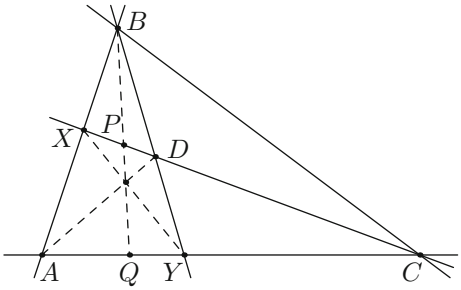
(1) Note that Alice cannot (directly) force Bob to choose some point on a line or on a circle, and this is often needed in the standard geometric constructions. But this is inessential since Alice can achieve this goal in several steps. First she asks to add points on both sides of the line or circle (selecting two small open sets on both sides in such a way that every interval with endpoints in these open sets intersects the line or circle), then asks to connect these points by a line, and then asks to add the intersection point of this new line and the original one.

(2) On the other hand, according to our rules, Alice can specify with arbitrarily high precision where the new point should be (by choosing a small open set). A weaker (for Alice) option would be to allow her to choose a connected component of the complement of the union of all objects in the current position. Then Bob should add some point of this component to the current position.

Proposition 1. *This restriction does not change the notion of a constructible point.*

Proof. Idea: Using the weaker option, Alice may force Bob to put enough points to make the set of derivable points dense, and then use the first three options to get a point in an arbitrary open set.

Let us explain the details. First, she asks for an arbitrary point A, then for a point B that differs from A, then for line AB, then for a point C outside line AB (thus having the triangle ABC), then for the sides of this triangle, and then for a point D inside the triangle. (All this is allowed in the restricted version.)



Now the points P and Q obtained as shown are derivable (after the projective transformation that moves B and C to infinity, the points P and Q become the midpoints of XD and AY). Repeating this construction, we get a dense set of derivable points on intervals XD and AY, then the dense set of derivable points in the quadrangle $AXDY$ and then in the entire plane.

Now, instead of asking Bob for a point in some open set U, Alice may force him to include one of the derivable points (from the dense set discussed above) that is in U.

This definition of constructibility turns out to be equivalent to the negative definition suggested by Akopyan and Fedorov [1]. They define non-constructibility as follows: an object x is *non-constructible* from a finite set E of objects if there exists a set $E' \supset E$ that is closed under the operations of adding points, lines, and circles (contains all objects derivable from E'), contains an everywhere dense set of points, but does not contain x.

Proposition 2 (Akopyan–Fedorov). *This negative definition is equivalent to the game-theoretic definition given above.*

Proof. The equivalence is essentially proven as [1, Proposition 15, p. 9], but Akopyan and Fedorov avoided stating explicitly the game-theoretic definition and spoke about "algorithms" instead (without an exact definition).

Assume that x is non-constructible from E according to the negative definition. Then Bob can prevent Alice from winning by always choosing points from E' when Alice asks for a point in an open set. Since E' is dense, these points are enough. If Bob follows this strategy, then the current position will always be a subset of E' and therefore will never contain x.

On the other hand, assume that x in not constructible from E in the sense of the positive definition. Consider the following strategy for Alice. She takes some triangle abc and point d inside it and ask Bob to add points A, B, C, D that belong to some small neighborhoods of a, b, c, d respectively. The size of these neighborhoods guarantees that ABC is a triangle and D is a point inside ABC. There are two cases:

- for every choice of Bob Alice has a winning strategy in the remaining game;
- there are some points A, B, C, D such that Alice does not have a winning strategy in the remaining game.

In the first case Alice has a winning strategy in the entire game and x is constructible. In the second case we consider the set E' of all objects derivable from $E \cup \{A, B, C, D\}$. As we have seen in the proof of the previous proposition, this set is dense. Therefore, x is non-constructible in the sense of the negative definition.

The advantage of the game definition is that it can be reasonably extended to the uniform case. For the uniform case the game is no more a full-information game. Alice sees only the names (and types) of geometric objects in E, and assigns names to new objects produced by Bob. One should agree also how Alice can get information about the configuration and how she can specify the connected component when asking Bob for a point in this component. For example, we may assume that Alice has access to the list of all connected components and the full topological information about the structure they form, as well as the places of objects from E in this structure. Then Alice may choose some component and request a point from it. To win, Alice needs to specify the name of the required object x. After we agree on the details of the game, we may define construction algorithms as computable strategies for such a game. (Note that in this version Alice deals only with finite objects).

6 Formal Definitions Are Important

In fact, the absence of formal definitions and exact statements is more dangerous than one could think. It turned out that some classical and well known arguments contain a serious gap that cannot be filled without changing the argument. This happened with a proof (attributed to Hilbert in [5]) that one cannot find the center of a given circle using only a straightedge. It is reproduced in many

popular books (see, e.g., [6,10,15]) and all the arguments (at least in the four sources mentioned above) have the same gap. They all go as follows [10, p. 18]:

Let the construction be performed in a plane P_1 and imaging a transformation or mapping T of the plane P_1 into another plane P_2 such that:

(a) straight lines in P_1 transform into straight lines in P_2 $\langle \ldots \rangle$
(b) The circumference C of our circle is transformed into a circumference $T(C)$ for some circle in P_2.

As the steps called for in the construction are being performed in P_1, they are being faithfully copied in P_2. Thus when the construction in P_1 terminates in the centre O of C, the "image" construction *must* terminate in the centre $T(O)$ of the circle $T(C)$.

Therefore if one can exhibit a transformation T satisfying (a) and (b), but such that $T(O)$ *is not* the centre of $T(C)$, then the impossibility of constructing the centre of a circle by ruler alone will be demonstrated.

Such a transformation indeed exists, but the argument in the last paragraph has a gap. If we understand the notion of construction in a non-uniform way and require that the point was among the points constructed, the argument does not work since the center of $T(C)$ could be the image of some other constructed point. If we use some kind of the uniform definition and allow tests, then these tests can give different results in P_1 and P_2 (the projective transformation used to map P_1 into P_2 does not preserve the ordering), so there is no reason to expect that the construction is "faithfully copied". And a uniform definition that does not allow tests and still is reasonable, is hard to imagine (and not given in the book). Note also that some lines that intersect in P_1, can become parallel in P_2.

It is easy to correct the argument and make it work for the definition of constructibility given above (using the fact that there are many projective mappings that preserve the circle), but still one can say without much exaggeration that the first correct proof of this impossibility result appeared only in [1]. One can add also that the stronger result about two circles that was claimed by Cauer [5] and reproduced with a similar proof in [15], turned out to be plainly false as shown in [1], and the problems in the proof were noted already by Gram [8]. It is not clear why Gram did not question the validity of the classical proof for one circle, since the argument is the same. Gram did not try to give a rigorous definition of the notion of a geometric construction, speaking instead about constructions in the "ordered plane" and referring to Bieberbach's book [3] that also has no formal definitions.

The weak version of Cauer's result saying that for some pairs of circles one cannot construct their centers, can be saved and proven for the definition of constructibility discussed above (see [1] and the popular exposition in [17]).

It would be interesting to reconsider the other results claimed about geometric constructions (for example, in [9,19–21]) to see whether the proofs work for some clearly defined notion of a geometric construction. Note that in some cases (e.g., for Tietze's results) some definition of the geometric construction for the uniform case is needed (and the negative definition is not enough).

Acknowledgements. The authors thank Sergey Markelov, Arseny Akopyan, Roman Fedorov and their colleagues at Moscow State University and LIRMM (Montpellier) for interesting discussions.

References

1. Akopyan, A., Fedorov, R.: Two circles and only a straightedge (2017). https://arxiv.org/abs/1709.02562
2. Baston, V.J., Bostock, F.A.: On the impossibility of ruler-only constructions. Proc. Am. Math. Soc. **110**(4), 1017–1025 (1990)
3. Bieberbach, L.: Theorie der Geometrischen Konstruktionen. Springer, Basel (1952). https://doi.org/10.1007/978-3-0348-6910-2
4. Borel, E.: Le calcul des intégrales définies. J. Math. pures appl. ser. 6 **8**(2), 159–210 (1912)
5. Cauer, D.: Über die Konstruktion des Mittelpunktes eines Kreises mit dem Lineal allein. Math. Annallen **73**, 90–94 (1913). A correction: **74**, 462–464
6. Courant, R., Robbins, H., revised by Stewart, I.: What is Mathematics? An Elementary Approach to Ideas and Methods. Oxford University Press, Oxford (1996)
7. Engeler, E.: Remarks on the theory of geometrical constructions. In: Barwise, J. (ed.) The Syntax and Semantics of Infinitary Languages. LNM, vol. 72, pp. 64–76. Springer, Heidelberg (1968). https://doi.org/10.1007/BFb0079682
8. Gram, C.: A remark on the construction of the centre of a circle by means of the ruler. Math. Scand. **4**, 157–160 (1956)
9. Hilbert, D.: The foundations of geometry, authorized translation by E.J. Townsend, Ph.D., University of Illinois (1902)
10. Kac, M., Ulam, S.M.: Mathematic and Logic. Dover publications, New York (1992)
11. Kijne, D.: Plane construction field theory. Ph.D. thesis, promotor H. Freudenthal, van Gorcum & Co., N.V., G.A. Hak, H.J. Prakke, 28 May 1956
12. Kutuzov, B.V.: Studies in mathematics, vol. IV, Geometry (trans. by L.I. Gordon, E.S. Shater) School Mathematics Study Group, Chicago (1960)
13. Manin, Y.: On the decidability of geometric construction problems using compass and straightedge [Russian]. Encyclopedia of Elementary Mathematics, Geometry, Moscow, vol. IV, pp. 205–227 (1963)
14. Martin, G.E.: Geometric Constructions. Springer, New York (1998). https://doi.org/10.1007/978-1-4612-0629-3
15. Rademacher, H., Toeplitz, O.: Von Zahlen und Figuren, 2nd edn. Springer, Heidelberg (1933). https://doi.org/10.1007/978-3-662-36239-6
16. Schreiber, P.: Theorie der Geometrischen Konstruktionen. VEB Deutscher Verlag der Wissenschaften, Berlin (1975)
17. Shen, A.: Hilbert's Error? (2018). https://arxiv.org/abs/1801.04742
18. Tao, T.: A geometric proof of the inpossibility of angle trisection. https://terrytao.wordpress.com/2011/08/10/a-geometric-proof-of-the-impossibility-of-angle-trisection-by-straightedge-and-compass/
19. Tietze, H.: Über dir Konstruierbarkeit mit Lineal und Zirkel, Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften, Abt. IIa, 735–757 (1909). https://www.biodiversitylibrary.org/item/93371
20. Tietze, H.: Über die mit Lineal und Zirkel und die mit dem rechten Zeichenwinkel lösbaren Konstruktionaufgaben I. Math. Zeitschrift **46**, 190–203 (1940). http://www.digizeitschriften.de/dms/img/?PID=GDZPPN002379074

21. Tietze, H.: Zur Analyse der Lineal- und Zirkelkonstruktionen. I. Sitzungsberichte der mathematisch-naturwissenschaftlichen Abteilung der Bayrischen Akademie der Wissenschaften zu München, 1944, Heft III, Sitzungen Oktober–Dezember, pp. 209–231, München (1947). http://publikationen.badw.de/003900992.pdf
22. Wantzel, M.L.: Recherches sur les moyens de reconnaître si un probléme de Géométrie peut se résoudre avec la règle et le compas. J. Math. pures Appl. 1re série **2**, 366–372 (1837)