Logic of Ruler and Compass Constructions

M. Beeson

San José State University

Abstract. We describe a theory **ECG** of "Euclidean constructive geometry". Things that **ECG** proves to exist can be constructed with ruler and compass. **ECG** permits us to make constructive distinctions between different forms of the parallel postulate. We show that Euclid's version, which says that under certain circumstances two lines meet (i.e., a point of intersection exists) is not constructively equivalent to the more modern version, which makes no existence assertion but only says there cannot be two parallels to a given line. Non-constructivity in geometry corresponds to case distinctions requiring different constructions in each case; constructivity requires continuous dependence on parameters. We give continuous constructions where Euclid and Descartes did not supply them, culminating in geometrical definitions of addition and multiplication that do not depend on case distinctions. This enables us to reduce models of geometry to ordered field theory, as is usual in non-constructive geometry. The models of ECG include the set of pairs of Turing's constructible real numbers [7].

1 Introduction

Euclid's geometry, written down about 300 BCE, has been extraordinarily influential in the development of mathematics, and prior to the twentieth century was regarded as a paradigmatic example of pure reasoning.

In this paper, we re-examine Euclidean geometry from the viewpoint of constructive mathematics. The phrase "constructive geometry" suggests, on the one hand, that "constructive" refers to geometrical constructions with straightedge and compass. On the other hand, the word "constructive" may suggest the use of intuitionistic logic. We investigate the connections between these two meanings of the word. Our method is to keep the focus on the body of mathematics in Euclid's *Elements*, and to examine what in Euclid is constructive, in the sense of "constructive mathematics". Our aim in the first phase of this research was to formulate a suitable formal theory that would be faithful to both the ideas of Euclid and the constructive approach of Errett Bishop [4]. We achieved this aim by formulating a theory **ECG** of "Euclidean constructive geometry", first presented in [2], but improved in [3].

In constructive mathematics, if one proves something exists, one has to show how to construct it. In Euclid's geometry, the means of construction are not arbitrary computer programs, but ruler and compass. Therefore it is natural to look for a theory that has function symbols for the basic ruler-and-compass

constructions. The terms of such a theory correspond to ruler-and-compass constructions. Our first main result is that when **ECG** proves that something exists, that something can be constructed with ruler and compass.

In number theory, if one proves an existence theorem, then for a constructive version, one has to show how to compute the desired number as a function of the parameters. In analysis, if one proves an existence theorem, one has to be able to compute approximations to the desired number from approximations to the parameters. In particular, the solution will depend continuously on parameters, at least locally. This feature of constructive analysis depends, in a way, on what we think it means "to be given" a number x. Whatever that may mean, it surely means that we have a way to get a rational approximation to x within any specified limit of accuracy. Geometry is more like analysis than number theory, in the sense that we do not want to assume in advance that points can be given to us all at once in a completely determined location; points are given only approximately, by dots on paper or a computer screen, or in Euclid's case, by indentations in sand (the Greeks drew their diagrams in sand). It might be doubtful whether two such points coincide; in such a case one would have to ask the one who made the diagram to refine it. It follows that in constructive geometry, we should have local continuous dependence of constructions on parameters. We can see that dramatically in computer animations of Euclidean constructions, in which one can select some of the original points and drag them, and the entire construction "follows along." One might formulate a program of "continuous geometry", in which one allows only constructions that depend continuously on parameters. It turns out that this is just another way of viewing constructive geometry, since theorems proved without non-constructive case distinctions will be implemented by continuous ruler-and-compass constructions. One line of research has thus been to identify and repair uses of non-constructive case distinctions. There are several important places where repair is needed, but it is possible. Thus the "**C**" in **ECG** could just as well be read as "continuous", instead of "constructive."

Once we have a good formal theory of constructive geometry, the possibility opens up to prove independence results. Our most striking results concern the different formulations of Euclid's parallel postulate. Euclid's original version (Euclid 5) is not the same as the version more commonly used today (Playfair's axiom). The difference is that Euclid 5 says that under certain conditions, two lines must meet, while Playfair's axiom says that there cannot be two different parallels to line L through point P. Thus Euclid 5 makes an existence assertion, but Playfair does not. We prove that Playfair does not imply Euclid 5 in **ECG** minus the parallel axiom of **ECG**.

In classical (i.e. nonconstructive) geometry, there are theorems that show that models of geometrical theories all have the form F^2, where F is an ordered field, and the geometrical relations of incidence and betweenness are defined as usual in analytic geometry. Different geometric axioms correspond to different axioms in ordered field theory. When the geometric axioms are those for ruler and compass, we get Euclidean fields (those in which positive elements have square roots). We show that this paradigm extends to constructive geometry as well.

This is not trivial, because we need to give geometrical definitions of addition and multiplication (of segments, or of points on a line) that are continuous in parameters, in particular, do not require case distinctions about the sign to construct the sum and product.

Once that is done, we move on to consider the models F^2, and we find that there are three different possible definitions of "constructive Euclidean field". The difference hinges on when the reciprocal $1/x$ is defined: either positive elements have reciprocals, or nonzero elements have reciprocals, or elements without reciprocals are zero. To prove two of these three versions equivalent requires either proof by contradiction or non-constructive case distinctions. Each of these versions corresponds to geometry, with a different version of the parallel axiom.

We obtain our independence proofs by constructing models of one kind of constructive ordered field theory that do not satisfy the next stronger kind. Of course these are not fields in the usual sense, because these field theories are non-constructively equivalent; they are what is known as Kripke models. Their construction involves taking the field elements to be certain real-valued functions, corresponding to points whose location is not yet "pinned down."

This short paper omits proofs, discussion of past axiomatizations of geometry, philosophical discussions, and detailed discussions of the relations between this work and the work of others. All these things can be found in [3]. Heartfelt thanks to Marvin Greenberg and Freek Wiedijk for their comments and suggestions.

2   Is Euclid's reasoning constructive?

Euclid's reasoning is generally constructive; indeed the only irreparably non-constructive proposition is Book I, Prop. 2, which shows that the rigid compass can be simulated by a collapsible compass. We just take Euclid I.2 as an axiom, thus requiring a rigid compass in **ECG**. Only one other repair is needed, in the formulation of the parallel axiom, as we shall see below. Euclid did not deal with disjunctions explicitly, and all his theorems are of the form: Given certain points related in certain ways, we can construct (zero or more) other points related to the given points and each other in certain ways. Euclid has been criticized (as far back as Geminus and Proclus) for ignoring case distinctions in a proof, giving a diagram and proof for only one case. Since case distinctions (on whether $ab = cd$ or not) are non-constructive, these omissions are *prima facie* non-constructive. However, these non-constructive proof steps are eliminable, as we will explain.

An example of such an argument in Euclid is Prop. I.6, whose proof begins

> Let ABC be a triangle having the angle ABC equal to the angle ACB. I say that the side AB is also equal to the side AC. For, if AB is unequal to AC, one of them is greater. Let AB be greater, ...

The same proof also uses an argument by contradiction in the form $\neg x \neq y \ \rightarrow \ x = y$. This principle, the "stability of equality", is an axiom of **ECG**, and is universally regarded as constructively acceptable. The conclusion of I.6, however,

is negative (has no ∃ or ∨), so we can simply put double negations in front of every step, and apply the stability of equality once at the end.

Prop. I.26 is another example of the use of the stability of equality: "…DE is not unequal to AB, and is therefore equal to it."

To put the matter more technically, in constructive logic we have $P \rightarrow \neg\neg P$, and although generally we do not have $\neg\neg P \rightarrow P$, we do have it for quantifier-free, disjunction-free P. We can double-negate $A \vee \neg A \rightarrow B$, obtaining $\neg\neg(A \vee \neg A) \rightarrow \neg\neg B$, and then the hypothesis is provable, so we have $\neg\neg B$, and hence B since B is quantifier-free and disjunction-free. The reason why this works throughout Euclid is that the *conclusions* of Euclid's theorems are all quantifier-free and disjunction-free. Euclid never even *thought* of stating a theorem with an "or" in it. The bottom line is that Euclid is constructive as it stands, except for Book I, Prop. 2, and the exact formulation of the parallel postulate. These problems are remedied in **ECG** by taking Book I, Prop. 2 as an axiom and strengthening the parallel postulate as discussed below. We also take as an axiom $\neg\neg\mathbf{B}(x,y,z) \rightarrow \mathbf{B}(x,y,z)$, or "Markov's principle for betweenness", enabling us to drop double negations on atomic sentences.

3 The elementary constructions

The Euclidean, or "elementary" constructions, are carried out by constructing lines and circles and marking certain intersection points as newly constructed points. The geometrical theory **ECG** given in [2] has terms to denote the geometrical constructions. These terms can sometimes be "undefined", e.g. if two lines are parallel, their intersection point is undefined. Therefore **ECG** is based on the logic of partial terms LPT [1], p. 97, in which there are atomic formulas $t \downarrow$ expressing that term t has a denotation ("is defined"). A model of such a theory can be regarded as a many-sorted algebra with partial functions representing the basic geometric constructions. Specifically, the sorts are *Point*, *Line*, and *Circle*. We have constants and variables of each sort.

ECG includes function symbols for the basic constructors and accessors, such as *Line*(A, B) for the line through A and B and *Circle*(A, B) for the circle through B with center A, and for the "elementary constructions" (each of which has type *Point*):

$$IntersectLines\,(Line\,K, Line\,L)$$
$$IntersectLineCircle1\,(Line\,L, Circle\,C)$$
$$IntersectLineCircle2\,(Line\,L, Circle\,C)$$
$$IntersectCircles1\,(Circle\,C, Circle\,K)$$
$$IntersectCircles2\,(Circle\,C, Circle\,K)$$

One can regard circles and lines as mere intermediaries; points are ultimately constructed from other points. (This was proved in [2].)

There is a second constructor for circles, which we can describe for short as "circle from center and radius", as opposed to the first constructor above,

"circle from center and point." Specifically $Circle3(A, B, C)$ constructs a circle of radius BC and center A, provided $B \neq C$. These two constructors for circles correspond to a "collapsible compass" and a "rigid compass" respectively. The compass of Euclid was a collapsible compass: you cannot use it to "hold" the length BC while you move one point of the compass to A. You can only use it to hold the radius AB while one point of the compass is fixed at A, so in that sense it corresponds to $Circle(A, B)$. The second constructor $Circle3$ corresponds to a rigid compass. The theory **ECG** includes $Circle3$, and in [2] we gave reasons why constructive geometry demands a rigid, rather than only a collapsible compass. In short, without a rigid compass, one cannot project a point P onto a line L, without making a case distinction between the case when P lies on L and the case when it does not; and the ability to make such projections is crucial to defining a coordinate system and showing how to perform addition and multiplication on segments.

There are three issues to decide:

– when there are two intersection points, which one is denoted by which term?
– In degenerate situations, such as $Line(P, P)$, what do we do?
– When the indicated lines and/or circles do not intersect, what do we do about the term(s) for their intersection point(s)?

Our answers to these questions are as follows. When the indicated lines or circles do not intersect, then the term for their intersection is "undefined". This can best be handled formally using the logic of partial terms, which we do in **ECG**; it can also be handled in other more cumbersome ways without modifying first-order logic. We take $Circle(P, P)$ to be defined, i.e., we allow circles of zero radius; that technicality makes the formal development smoother and seems philosophically unobjectionable–we just allow the two points of the compass to coincide. The point here is not so much that circles of zero radius are of interest, but that we do not want to force a case distinction as to whether the two points of the compass are, or are not, coincident. We take the two points of intersection of a line $Line(A, B)$ and a circle to occur in the same order as A and B occur on L. That means that lines are treated as having direction. Not only do they have direction, they "come equipped" with two points from which they were constructed. There are function symbols to recover those points from a line. $Line(P, P)$ is undefined, since having it defined would destroy continuous dependence of $Line(P, Q)$ on P and Q.

The two intersection points $p = IntersectCircles1(C, K)$ and $q = IntersectCircles2(C, K)$ are to be distinguished as follows: With a the center of C and b the center of K we should have abp a right turn, and abq a left turn. But can "right turn" and "left turn" be defined? What we do is to *define Right* and *Left* using equations involving $IntersectCircles1$ and $IntersectCircles2$; then we give axioms about $Right$ and $Left$, namely that if abc is a left turn, then c and d are on the same side of $Line(a, b)$ if and only if abd is a left turn, and c and d are on opposite sides of $Line(a, b)$ if and only if abd is a right turn. Note that neither this issue nor its solution has to do with constructivity, but simply with the introduction of function symbols corresponding to the elementary constructions.

4 Models of ruler-and-compass geometry

There are several interesting models of **ECG**, even with classical logic, of which we now mention four. The *standard plane* is $\mathbb{R}^2$, with the usual interpretation of points, lines, and planes. The *Turing plane* has for its points pairs of computable real numbers [7]. The *algebraic plane* has for its points pairs of algebraic numbers. The *Tarski plane* has for points just those points constructible with ruler and compass; this is $\mathbb{K}^2$ where $\mathbb{K}$ is the least subfield of the reals closed under square roots of positive elements. The theory **ECG** uses the same primitive relations as Tarski and Hilbert: betweenness $\mathbf{B}(a, b, c)$ and equidistance $ab = cd$. Hilbert used strict betweenness and Tarski allowed $\mathbf{B}(x, x, x)$; we follow Hilbert.

5 Three versions of the parallel postulate

Let P be a point not on line L. We consider lines through P that do not meet L (i.e., are parallel to L). Playfair's version of the parallel postulate says that two parallels to L through P are equal. Recall that Euclid's postulate 5 is

> *If a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles.*

Fig. 1. Euclid 5: M and L must meet on the right side, provided $\mathbf{B}(q, a, r)$ and pq makes alternate interior angles equal with K and L. The point at the open circle is asserted to exist.
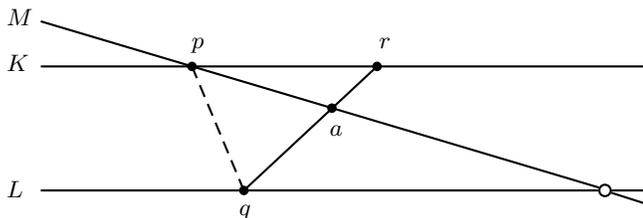
But this formulation of Euclid 5 makes use of the notion of "alternate interior angles", while angles are not directly treated in **ECG**, but instead are treated as triples of points. A version of Euclid 5 that does not mention angles is given in Fig. 2.

Although we have finally arrived at a satisfactory formulation of Euclid 5, that formulation is satisfactory only in the sense that it accurately expresses what Euclid said. It turns out that this axiom is not satisfactory as a parallel postulate for **ECG**. The main reason is that it is inadequate to define division geometrically. Here is why: As x gets nearer and nearer to 0, the number $1/x$ requires a line of smaller and smaller slope to meet a certain horizontal line. If x passes through zero, this intersection point "goes to infinity", then is undefined when $x = 0$, but then "reappears on the other side", coming in from minus

Fig. 2. Euclid 5: M and L must meet on the right side, provided $\mathbf{B}(q,a,r)$ and $pt = qt$ and $rt = st$.
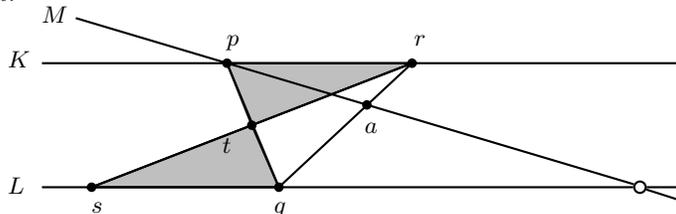

infinity. Without knowing the sign of x, we will not know on which side of the transversal pq the two adjacent interior angles will make less than two right angles. In other words, with Euclid 5, we will only be able to divide by a number whose sign we know; and the principle $x \neq 0 \;\rightarrow\; x < 0 \vee x > 0$ is not an axiom (or theorem) of **ECG**. The conclusion is that if we want to divide by nonzero numbers, we need to strengthen Euclid's parallel axiom.

We make three changes in Euclid 5 to get the "strong parallel postulate":

(i) We change the hypothesis $\mathbf{B}(q,a,r)$ to $\neg on(a,K)$. In other words, we require that the two adjacent interior angles do not make exactly two right angles, instead of requiring that they make less than two right angles.

(ii) We change the conclusion to state only that M meets L, without specifying on which side of the transversal pq the intersection lies.

(iii) We drop the hypothesis $\neg\, on\,(p,L)$.

Fig. 3. Strong Parallel Postulate: M and L must meet (somewhere) provided a is not on K and $pt = qt$ and $rt = st$.
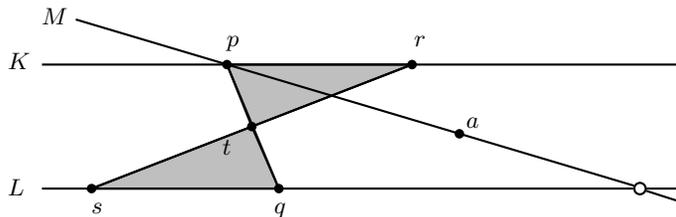
The strong parallel axiom differs from Euclid's version in that we are not required to know in what *direction* M passes through P; but also the conclusion is weaker, in that it does not specify *where* M must meet L. In other words, the betweenness hypothesis of Euclid 5 is removed, and so is the betweenness conclusion. Since both the hypothesis and conclusion have been changed, it is not immediate whether this new postulate is stronger than Euclid 5, or equivalent, or possibly even weaker, but it turns out to be stronger–hence the name.

6   Constructive Geometry and Euclidean Fields

Classical Euclidean geometry has models $K^2 = K \times K$ where K is a Euclidean field, i.e. an ordered field in which nonnegative elements have square roots. We

take that definition also constructively, and we define a *Euclidean* ring to be an ordered ring in which nonnegative elements have square roots. We use a language with symbols $+$ for addition and $\cdot$ for multiplication, and a unary predicate $P(x)$ for "x is positive". A *Euclidean field* is a Euclidean ring in which nonzero elements have reciprocals. Constructively, we also need two weaker notions: Euclidean rings in which positive elements have reciprocals, and Euclidean rings in which elements without reciprocals are zero and if x is greater than a positive invertible element, then x is invertible. These we call *Playfair rings*.

In order to show that the models of some geometrical theory T have the form F^2, one has to define addition and multiplication (of segments or points on a line) within T. This was first done by Descartes in [5], and again (in a different way) by Hilbert in [6]. These constructions, however, involve a non-constructive case distinction on the sign of the numbers being added or multiplied. To repair this problem requires rather more elaborate constructions, and to make those elaborate constructions work, one needs some more elementary case constructions to also work without case distinctions, for example, constructing a line through a point P perpendicular to a line L, without a case distinction as to whether P is or is not on L. These problems are solved, and the (rather lengthy) solutions are presented in detail, in [3].

We can prove that the models of **ECG** are of the form F^2, where F is a Euclidean field. More specifically, given such a field, we can define betweenness, incidence, and equidistance by analytic geometry and verify the axioms of **ECG**. Conversely, and this is the hard part, we can define multiplication, addition, and division of points on a line (having chosen one point as zero), in **ECG**. It turns out that we need the strong parallel axiom to do that. If we replace the parallel axiom of **ECG** by Euclid's parallel postulate, we get instead models of the form F^2, where F is a Euclidean ring in which nonzero elements have reciprocals, but we cannot go the other way by defining multiplication and addition geometrically without the strong parallel axiom. (That is, if we only had Euclid 5, we would need case distinctions, as Hilbert and Descartes did.)

We now work out the field-theoretic version of Playfair's axiom. Playfair says, if P is not on L and K is parallel to L through P, that if line M through P does not meet L then $M = K$. Since $\neg\neg M = K \ \to \ M = K$, Playfair is just the contrapositive of the parallel axiom of **ECG**, which says that if $M \neq K$ then M meets L. Hence it corresponds to the contrapositive of $x \neq 0 \ \to \ 1/x \downarrow$; that contrapositive says that if x has no multiplicative inverse, then $x = 0$. Thus Playfair geometries have models F^2 where F is a Playfair ring (as defined above). (We cannot prove the converse because we need the strong parallel axiom to verify multiplication and addition).

7 What ECG proves to exist, can be constructed with ruler and compass

In [2], we proved that if **ECG** proves an existential statement $\exists y A(x, y)$, then there is a term t of **ECG** such that **ECG** proves $A(x, t(x))$. In words: things that

ECG can prove to exist, can be constructed with ruler and compass. Of course, the converse is immediate: things that can be constructed with ruler and compass can be proved to exist in **ECG**. Hence the two meanings of "constructive" coincide for **ECG**: it could mean "proved to exist with intuitionistic logic" or it could mean "constructed with ruler and compass."

The technique of the proof is to apply Gentzen's cut-elimination theorem. What makes it applicable is that the axiomatization of **ECG** has two important properties: it is *quantifier-free*, and it is *disjunction-free*. It was not difficult to axiomatize **ECG** in this way–we just followed Euclid. In [3] we draw on Tarski's approach to achieve a short elegant list of axioms, but that is not essential to the analysis of the parallel axiom. There are many ways to axiomatize geometry.

8   Independence results for the Parallel Axioms

The reduction of geometry to field theory described above shows that (relative to a base theory), the strong parallel axiom implies Euclid's postulate 5 (since if reciprocals of non-zero elements exist, then of course reciprocals of positive elements exist). (A direct proof is in [3].) And Euclid 5 easily implies Playfair's postulate. Our main theorem is that neither of these two implications can be reversed.

Theorem 1. *Euclid 5 does not imply the strong parallel axiom, and Playfair does not imply Euclid 5, in* **ECG** *minus its parallel axiom.*

Proof sketch of the first claim (A detailed proof can be found in [3]). Since non-constructively, the implications *are* reversible, we cannot hope to give counterexamples. In terms of field theory, we won't be able to construct a Euclidean ring in which positive elements have reciprocals but nonzero elements do not. The proof proceeds by constructing appropriate Kripke models. To show that Euclid 5 does not prove the strong parallel axiom, it suffices to prove the corresponding result in ordered field theory: the axiom that positive elements have reciprocals does not imply that all nonzero elements have reciprocals. That does suffice, in spite of the fact that we have full equivalence between geometry and field theory only for **ECG** and Euclidean fields, for if the weaker geometry proved the strong parallel axiom SP, then the interpretation of SP in field theory would be provable, as *that* direction does work, and the interpretation of SP implies that nonzero elements have reciprocals.

Then we need a Kripke model in which positive elements have reciprocals, but nonzero elements do not necessarily have reciprocals. We construct such a Kripke model whose "points" are functions from $\mathbb{R}$ to $\mathbb{R}$. The function f is *positive semidefinite* if $f(x) \geq 0$ for all real x. Let $\mathbb{K}$ be the least subfield of the reals closed under square roots of positive elements. Let $\mathcal{A}$ be the least ring of real-valued functions containing polynomials with coefficients in $\mathbb{K}$, and closed under reciprocals and square roots of positive semidefinite functions. For example

$$\sqrt{\sqrt{1+t^2} + \sqrt{1+t^4}} + \frac{1}{1+t^2}$$

is in $\mathcal{A}$, but $1/t$ is not in $\mathcal{A}$. We take $\mathcal{A}$ as the root of a Kripke model, interpreting the positivity predicate $P(x)$ to mean x is positive definite. We show (using Pusieux series) that each member of $\mathcal{A}$ has finitely many zeroes and singularities and that there is a countable set Ω including all zeroes and singularities, whose complement is dense in $\mathbb{R}$. For $\alpha \notin \Omega$, we define $\mathcal{A}_\alpha$ by interpreting $P(x)$ to hold if and only if $x(\alpha) > 0$. In our Kripke model, $\mathcal{A}_\alpha$ lies immediately above the root. Now t is a nonzero element without a reciprocal. But if x is positive, then $x(\alpha) > 0$ for all $\alpha \notin \Omega$, and since the complement of Ω is dense and x is continuous, x is positive semidefinite, so $1/x$ exists in $\mathcal{A}$.

9 Conclusions

Euclid needs only two modifications to be completely constructive: we have to postulate a rigid compass, rather than relying on Prop. I.2 to simulate it, and we have to take the strong parallel axiom instead of Euclid 5. With those changes Euclid is entirely constructive, and **ECG** formalizes Euclid nicely. **ECG** has the nice property that things it can prove to exist can be constructed with ruler and compass, and it permits us to distinguish between versions of the parallel axiom with different constructive content, even though non-constructively they are equivalent. The classical constructions used to give geometrical definitions of addition and multiplication involve non-constructive case distinctions, but these can be replaced by more elaborate constructions that are continuous (and constructive), so geometry can still be shown equivalent to the theory of Euclidean fields, and different versions of the parallel axiom correspond to weakenings of the field axiom about reciprocals. We can then use Kripke models whose points are certain real-valued functions to establish formal independence results about the different versions of the parallel axiom.

References

1. Beeson, M.: Foundations of Constructive Mathematics. Springer-Verlag, Berlin Heidelberg New York (1985)
2. Beeson, M.: Constructive geometry. In: Arai, T. (ed.) Proceedings of the Tenth Asian Logic Colloquium, Kobe, Japan, 2008. pp. 19–84. World Scientific, Singapore (2009)
3. Beeson, M.: Foundations of Constructive Geometry (available on the author's website, `wwwmichaelbeesoncom/research/papers/pubshtml`)
4. Bishop, E.: Foundations of Constructive Analysis. McGraw-Hill, New York (1967)
5. Descartes, R.: The Geometry of Rene Descartes, with a facsimile of the first edition, published in 1637 as an appendix to *Discours de la Methode*. Dover, New York (1952), original title, *La Geometrie*. Facsimile with English translation by David Eugene Smith and Marcia L. Latham
6. Hilbert, D.: Foundations of Geometry (Grundlagen der Geometrie). Open Court, La Salle, Illinois (1960), second English edition, translated from the tenth German edition by Leo Unger. Original publication date, 1899.
7. Turing, A.: On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society, Series 2 42 (1936-7)