Analytic geometry

David Pierce

December 5, 2013, 2:03 p.m.

Mathematics Department
Mimar Sinan Fine Arts University
dpierce@msgsu.edu.tr
http://mat.msgsu.edu.tr/~dpierce/

Contents

List of Figures

Introduction

The writing of this report was originally provoked both by frustration with the lack of rigor in analytic geometry texts and by a belief that this problem can be remedied by attention to mathematicians like Euclid and Descartes, who are the original sources of our collective understanding of geometry. Analytic geometry arose with the importing of algebraic notions and notations into geometry. Descartes, at least, justified the algebra geometrically. Now it is possible to go the other way, using algebra to justify geometry. Textbook writers of recent times do not make it clear which way they are going. This makes it impossible for a student of analytic geometry to get a correct sense of what a *proof* is. I find this unacceptable.

If it be said that analytic geometry is not concerned with proof, I would respond that in this case the subject pushes the student back to a time before Euclid, but armed with many more unexamined presuppositions. Students today have the idea that every line segment has a length, which is a positive real number of units, and conversely every positive real number is the length of some line segment. The latter presupposition is quite astounding, since the real numbers compose an uncountable set. Euclidean geometry can in fact be done in a countable space, as David Hilbert points out.

I made notes on some of these matters. The notes grew into this report as I found more and more things that seemed worth saying. There are still many more avenues to explore. Some of the notes here are just indications of what can be investigated further, either in mathematics itself or in the existing literature about it. Meanwhile, the contents of the numbered chapters of this report might be summarized as follows.

1. The logical foundations of analytic geometry as it is often taught are unclear. The subject is not presented rigorously.
2. What is rigor anyway? I consider some modern examples

where rigor is lacking. Rigor is not an absolute notion, but must be defined in terms of the audience being addressed.

3. Ancient mathematicians like Euclid and Archimedes still set the standard for rigor.
4. I have suggested *how* they are rigorous; but *why* are they rigorous? I don't know. But we still expect rigor from our students, if only because we expect them to be able to justify their answers to the problems that we assign to them—or if we don't expect this, we ought to.
5. I look at an old analytic geometry textbook that I learned something from as a child, but that I now find mathematically sloppy.
6. Because that book uses the odd terms *abscissa* and *ordinate* without explaining their origin, I provide an explanation, which involves the conic sections as presented by Apollonius.
7. The material on conic sections spills over to another chapter. Here we start to look in more detail at the geometry of Descartes. How Apollonius himself works out his theorems remains mysterious: Descartes's methods do not seem to illuminate those theorems.
8. I look at an analytic geometry textbook that I once taught from. It is more sophisticated than the textbook from my childhood. This makes its failures of rigor more frustrating to me and possibly more dangerous for the student.
9. I spell out more details of the justification of the use of algebra in geometry. Descartes acknowledges the need to provide this justification.
10. Finally, I review how the algebra of certain ordered fields can be used to obtain a Euclidean plane.

My scope here is the whole history of mathematics. Obviously I cannot give this a thorough treatment. I am not prepared to *try* to do this. To come to some understanding of a mathematician, one must *read* him or her; but I think one must read with a sense of what it means to do mathematics, *and* with an awareness that this sense may well differ from that of the mathematician whom one is reading. This awareness requires experience, in addition to the mere will to have it.

I have been fortunate to read old mathematics, both as a student and as a teacher, in classrooms where *everybody* is working through this mathematics and presenting it to the class. I am currently in the third year of seeing how new undergraduate mathematics students respond to Book I of Euclid's *Elements.* I continue to be surprised by what the students have to say. Mostly what I learn from the students themselves is how strange the notion of proof can be to some of them. This impresses on me how amazing it is that the *Elements* was produced in the first place. It may remind me that what Euclid even *means* by a proof may be quite different from what we mean today.

But the students alone may not be able to impress on me some things. Some students are given to writing down assertions whose correctness has not been established. These students' proofs may end up as a sequence of statements whose logical interconnections are unclear.[1] This is not the case with Euclid. And yet Euclid does begin each of his propositions with a bare assertion. He does not preface this enunciation or *protasis* (πρότασις) with the word "theorem" or "problem" as we might today. He does not have the typographical means that Heiberg uses in his own edition of Euclid to distinguish the protasis from the rest of the proposition. No, the protasis just sits there, not even preceded by the "I say that" (λέγω ὅτι) that may be seen further down in the proof. For me to notice this, naïve students were apparently not enough, but I had also to read Fowler's *Mathematics of Plato's Academy* [14, 10.4(e), pp. 385–6].

[1] Unfortunately some established mathematicians use the same style in their own lectures.

1 The problem

Textbooks of analytic geometry do not make their logical foundations clear. Of course I can speak only of the books that I have been able to consult: these are from the last century or so. Descartes's original presentation [8] in the 17th century *is* clear enough. In an abstract sense, Descartes may be no more rigorous than his successors. He does get credit for actually inventing his subject, or at least for introducing the notation we use today: minuscule letters for lengths, with letters from the beginning of the alphabet used for known lengths, and letters from the end for unknown lengths. As for his mathematics itself, Descartes explicitly bases it on an ancient tradition that culminates in the 4th century with Pappus of Alexandria.

More recent analytic geometry books start in the middle of things, but they do not make it clear what those things are. I think this is a problem. The chief aim of these notes is to identify this problem and its solution.

How can analytic geometry be presented rigorously? Rigor is not a fixed standard, but depends on the audience. Still, it puts some requirements on any work of mathematics, as I shall discuss in Chapter 2. In my own mathematics department, students of analytic geometry have had a semester of calculus, and a semester of synthetic geometry from its own original source, Book I of Euclid's *Elements* [13, 12]. They are the audience that I especially have in mind in my considerations of rigor. But I would suggest that any students of analytic geometry ought to come to the subject similarly prepared, at least on the geometric side.

Plane analytic geometry can be seen as the study of the Euclidean plane with the aid of a sort of rectangular grid that can be laid over the plane as desired. Alternatively, the subject can be seen as a discovery of geometric properties in the set of ordered pairs of real numbers. I propose to call these two approaches the *geometric* and the *algebraic,* respectively. Either approach can be made rigorous. But a course ought to be clear *which* approach is being taken.

Probably most courses of analytic geometry take the geometric approach, relying on students to know something of synthetic geometry already. Then the so-called Distance Formula can be justified by appeal to the Pythagorean Theorem. However, even in such a course, students might be asked to use algebraic methods to prove, for example, that *the base angles of an isosceles triangle are equal.*[1] Perhaps what would be expected is something like the following.

Proof 1. Suppose the vertices of a triangle are $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$, and the angles at $\boldsymbol{b}$ and $\boldsymbol{c}$ are β and γ respectively as in Figure 1.1. These

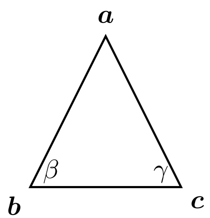


Figure 1.1: An isosceles triangle in a vector space

angles are given by

$$(\boldsymbol{a}-\boldsymbol{b})\cdot(\boldsymbol{c}-\boldsymbol{b}) = |\boldsymbol{a}-\boldsymbol{b}|\cdot|\boldsymbol{c}-\boldsymbol{b}|\cdot\cos\beta,$$
$$(\boldsymbol{a}-\boldsymbol{c})\cdot(\boldsymbol{b}-\boldsymbol{c}) = |\boldsymbol{a}-\boldsymbol{c}|\cdot|\boldsymbol{b}-\boldsymbol{c}|\cdot\cos\gamma,$$

We assume the triangle is isosceles, and in particular

$$|\boldsymbol{a}-\boldsymbol{b}| = |\boldsymbol{a}-\boldsymbol{c}|.$$

[1] I had a memory that this problem was assigned in an analytic geometry course that I was once involved with. However, I cannot find the problem in my files. I do find similar problems, such as to prove that the line segment bisecting two sides of triangle is parallel to the third side and is half its length, or to prove that, in an isosceles triangle, the median drawn to the third side is just its perpendicular bisector. In each case, the student is explicitly required to use analytic methods.

Then

$$\begin{aligned}(\boldsymbol{a}-\boldsymbol{c})\cdot(\boldsymbol{b}-\boldsymbol{c}) &= (\boldsymbol{a}-\boldsymbol{c})\cdot(\boldsymbol{b}-\boldsymbol{a}+\boldsymbol{a}-\boldsymbol{c})\\ &= (\boldsymbol{a}-\boldsymbol{c})\cdot(\boldsymbol{b}-\boldsymbol{a})+(\boldsymbol{a}-\boldsymbol{c})\cdot(\boldsymbol{a}-\boldsymbol{c})\\ &= (\boldsymbol{a}-\boldsymbol{c})\cdot(\boldsymbol{b}-\boldsymbol{a})+(\boldsymbol{a}-\boldsymbol{b})\cdot(\boldsymbol{a}-\boldsymbol{b})\\ &= (\boldsymbol{c}-\boldsymbol{a})\cdot(\boldsymbol{a}-\boldsymbol{b})+(\boldsymbol{a}-\boldsymbol{b})\cdot(\boldsymbol{a}-\boldsymbol{b})\\ &= (\boldsymbol{c}-\boldsymbol{b})\cdot(\boldsymbol{a}-\boldsymbol{b}),\end{aligned}$$

and so $\cos\beta = \cos\gamma$. □

If one has the Law of Cosines, the argument is simpler:

Proof 2. Suppose the vertices of a triangle are $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$, and the angles at $\boldsymbol{b}$ and $\boldsymbol{c}$ are β and γ respectively, again as in Figure 1.1. By the Law of Cosines,

$$\begin{gathered}|\boldsymbol{a}-\boldsymbol{c}|^2 = |\boldsymbol{a}-\boldsymbol{b}|^2 + |\boldsymbol{c}-\boldsymbol{b}|^2 - 2\cdot|\boldsymbol{a}-\boldsymbol{b}|\cdot|\boldsymbol{c}-\boldsymbol{b}|\cdot\cos\beta,\\ \cos\beta = \frac{|\boldsymbol{c}-\boldsymbol{b}|}{2\cdot|\boldsymbol{a}-\boldsymbol{b}|},\end{gathered}$$

and similarly

$$\cos\gamma = \frac{|\boldsymbol{b}-\boldsymbol{c}|}{2\cdot|\boldsymbol{a}-\boldsymbol{c}|}.$$

If $|\boldsymbol{a}-\boldsymbol{b}| = |\boldsymbol{a}-\boldsymbol{c}|$, then $\cos\beta = \cos\gamma$, so $\beta = \gamma$. □

In this last argument though, the vector notation is a needless complication. We can streamline things as follows.

Proof 3. In a triangle ABC, let the sides opposite A, B, and C have lengths a, b, and c respectively, and let the angles as B and C be β and γ respectively, as in Figure 1.2. If $b = c$, then

$$\begin{gathered}b^2 = c^2 + a^2 - 2ca\cos\beta,\\ \cos\beta = \frac{a}{2c} = \frac{a}{2b} = \cos\gamma.\end{gathered}$$ □

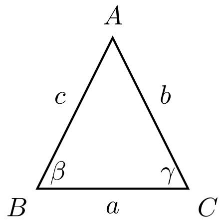


Figure 1.2: An isosceles triangle

Possibly this is not considered *analytic* geometry though, since coordinates are not used, even implicitly. We can use coordinates explicitly, laying down our grid conveniently:

Proof 4. Suppose a triangle has vertices $(0, a)$, $(b, 0)$, and $(c, 0)$, as in Figure 1.3. We assume $a^2 + b^2 = a^2 + c^2$, and so $b = -c$. In this

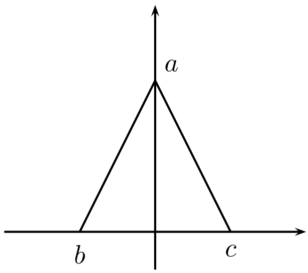


Figure 1.3: An isosceles triangle in a coordinate plane

case the cosines of the angles at $(b, 0)$ and $(c, 0)$ must be the same, namely $|b|/\sqrt{a^2 + b^2}$. □

In any case, as a proof of what is actually Euclid's Proposition I.5, this whole exercise is logically worthless, assuming we have taken the geometric approach to analytic geometry. By this approach, we shall have had to show how to erect perpendiculars to given straight lines, as in Euclid's Proposition I.11, whose proof relies ultimately on I.5.

One could perhaps develop analytic geometry on Euclidean principles without proving Euclid's I.5 as an independent proposition. For, the equality of angles that it establishes can be established immediately by means of I.4, the Side-Angle-Side theorem, by the method attributed to Pappus by Proclus [34, pp. 249–50]. In this case, one can still see clearly that I.4 is true, without needing to resort to any of the analytic methods suggested above.

2 Failures of rigor

The root meaning of the word *rigor* is stiffness. Rigor in a piece of mathematics is what makes it able to stand up to questioning. Rigor in mathematics *education* requires helping students to see what questions might be asked about a piece of mathematics.

An education in mathematics will take the student through several passes over the same subjects. With each pass, the student's understanding should deepen.[1] At an early stage, the student need not and cannot be told all of the questions that might be raised at a later stage. But if the mathematics of an early course resembles that of a different later course, it ought to be equally rigorous. Otherwise the older student might assume, wrongly, that the earlier mathematics could in fact stand up to the same scrutiny that the later mathematics stands up to. Concepts in an earlier course must not be presented in such a way that they will be misunderstood in a later course.

By this standard, students of calculus need not master the epsilon-delta definition of limit. If the students later take an analysis course, then they will fill in the logical gaps from the calculus course. The students are not going to think that everything was already proved in calculus class, so that epsilons and deltas are a needless complication. They may think there is no *reason* to prove everything, but that is another matter. If students of calculus never study analysis, but become engineers perhaps, or teachers of school mathematics, they are not likely to have false beliefs about what theorems can be proved in mathematics; they just will not have a highly developed notion of

[1] It might be counted as a defect in my own education that I did not have undergraduate courses in algebra and topology before taking graduate versions of these courses. Graduate analysis was for me a continuation of my high school course, which had been based on Spivak's *Calculus* [40] and (in small part) Apostol's *Mathematical Analysis* [4].

proof.

By introducing and using the epsilon-delta definition of limit at the very beginning of calculus, the teacher might actually violate the requirements of rigor, if he or she instills the false notion that there is no rigorous alternative definition of limits. How many calculus teachers, ignorant of Robinson's "nonstandard" analysis [36], will try to give their students some notion of epsilons and deltas, out of a misguided notion of rigor, when the intuitive approach by means of infinitesimals can be given full logical justification?

On the other hand, in mathematical circles, I have encountered disbelief that the real numbers constitute the unique complete ordered field. Since every valued field has a completion, it is possible to suppose wrongly that every ordered field has a completion. This confusion might be due to a lack of rigor in education, somewhere along the way.[2] There are (at least) two ways to obtain $\mathbb{R}$ from $\mathbb{Q}$. One way is to take the quotient of the ring of Cauchy sequences of rational numbers by the ideal of such sequences that converge to 0. Another way is to complete the ordering of $\mathbb{Q}$, as by taking Dedekind cuts, and then to show that this completion can be made into a field. The first construction can be generalized to apply to non-Archimedean ordered fields; the second cannot. More precisely, the second construction can be applied to a non-Archimedean field, but the result is not a field. It might be better to say that while the first construction achieves the completeness of $\mathbb{R}$ as a metric space, the second achieves the *continuity* of $\mathbb{R}$ as an ordered field. At any rate, continuity is the word Dedekind uses for what his construction is supposed to accomplish [7].

In an elementary course, the student may learn a theorem according to which certain conditions on certain structures are logically equivalent. But the theorem may use assumptions that are not spelled out. This is a failure of rigor. In later courses, the student learns logical equivalences whose assumptions *are* spelled out. The student may then assume that the earlier theorem is like the later

[2] I may have been saved from this confusion by Spivak's final chapter, "Uniqueness of the real numbers" [40, ch. 29].

ones. It may not be, and failure to appreciate this may cause the student to overlook some lovely pieces of mathematics. The word *student* here may encompass all of us.

The supposed theorem that I have in mind is that, in number theory, the principles of induction and well ordering are equivalent.[3] Proofs of two implications may be offered to back up this claim, though one of the proofs may be left as an exercise. The proofs will be of the standard form. They will look like other proofs. And yet, strictly speaking, they will make no logical sense, because:

- Induction is a property of algebraic structures in a signature with a constant, such as 1, and a singulary function symbol such as $'$ ("prime") for the operation of adding 1.
- Well ordering is a property of ordered structures.

When well ordering is used to prove induction, a set A is taken that contains 1 and is closed under adding 1, and it is shown that the complement of A cannot have a least element. For, the least element cannot be 1, and if the least element is $n+1$, then $n \in A$, so $n+1 \in A$, contradicting $n+1 \notin A$. It is assumed here that $n < n+1$. The correct conclusion is not that the complement of A is empty, but that *if* it is not, then its least element is not 1 and is not obtained by adding 1 to anything. Thus what is proved is the following.

Theorem 1. *Suppose $(S, 1, <)$ is a well-ordered set with least element 1 and with no greatest element, so that S is also equipped with the operation $'$ given by*

$$n' = \min\{x \colon n < x\}.$$

If

$$\forall x\ \exists y\ (x = 1 \lor y' = x), \tag{$*$}$$

then $(S, 1, ')$ admits induction.

[3]"Either principle may be considered as a basic assumption about the natural numbers" [36, ch. 2, p. 23]. I use this book as an example because it is otherwise so admirable.

The condition $(*)$ is not redundant. Every *ordinal number* in von Neumann's definition is a well-ordered set,[4] and every limit ordinal is closed under the operation $'$, but only the least limit ordinal, which is $\{0, 1, 2, \dots\}$ or ω, admits induction in the sense we are discussing.[5] The next limit ordinal, which is $\omega \cup \{\omega, \omega+1, \omega+2, \dots\}$ or $\omega+\omega$, does not admit induction; neither do any of the rest, for the same reason: they are not ω, but they include it. Being well-ordered is equivalent to admitting *transfinite* induction, but that is something else.

Under the assumption $n < n+1$, ordinary induction does imply well ordering. That is, we have the following.

Theorem 2. *Suppose* $(S, 1, ', <)$ *admits induction, is linearly ordered, and satisfies*

$$\forall x\ x < x'. \tag{†}$$

Then $(S, <)$ *is well ordered.*

However, the following argument is inadequate.

Standard proof. If a subset A of S has no least element, we can let B be the set of all n in S such that no element of $\{x \colon x < n\}$ belongs to A. We have $1 \in B$, since no element of the empty set belongs to A. If $n \in B$, then $n \notin A$, since otherwise it would be the least element of A; so $n' \in B$. By induction, B contains everything in S, and so A contains nothing. $\square$

We have tacitly used:

Lemma 1. *Under the conditions of the theorem,*

$$\forall x\ 1 \leqslant x.$$

[4] An ordinal in von Neumann's definition is a set, rather than the isomorphism-class of well-ordered sets that it was understood to be earlier. Von Neumann's original paper from 1923 is [43]. One can read it, but one must allow for some differences in notation from what is customary now. This is one difficulty of relying on special notation to express mathematics.

[5] The least element of ω is usually denoted not by 1 but by 0, because it is the empty set.

This is easily proved by induction. Trivially $1 \leqslant 1$. Moreover, if $1 \leqslant x$, then $1 < x'$, since $x < x'$ (and orderings are by definition transitive). But the standard proof of Theorem 2 also uses

$$\{x \colon x < n'\} = \{x \colon x < n\} \cup \{n\},$$

that is,

$$x < n' \Leftrightarrow x \leqslant n.$$

The reverse implication is immediate; the forward is the following.

Lemma 2. *Under the conditions of the theorem,*

$$\forall x \, \forall y \, (x < y' \Rightarrow x \leqslant y).$$

This is not so easy to establish, although there are a couple of ways to do it. The first method assumes we have established the standard properties of the set $\mathbb{N}$ of natural numbers, perhaps by using the full complement of Peano Axioms as in Landau [26].

Proof 1. By *recursion,* we define a homomorphism h from $(\mathbb{N}, 1, ')$ to $(S, 1, ')$. By induction in $\mathbb{N}$, h is order-preserving and therefore injective. By induction in S, h is surjective. Thus h is an isomorphism. Since $(\mathbb{N}, <)$ has the desired property, so does $(S, <)$. □

The foregoing proof can serve by itself as a proof of Theorem 2. An alternative, *direct* proof of Lemma 2 is as follows; I do not know a simpler argument.

Proof 2. We name two formulas,

$$\varphi(x, y) \colon x < y \Rightarrow x' \leqslant y,$$
$$\psi(x, y) \colon x < y' \Rightarrow x \leqslant y.$$

So we want to prove $\forall x \, \exists y \, \psi(x, y)$. Since $<$ is a linear ordering, we have

$$(\ddagger) \qquad \varphi(x, y) \Leftrightarrow \psi(y, x).$$

We can prove also, as a lemma,

$$(\S) \qquad \varphi(x,y) \land \psi(x,y) \Rightarrow \varphi(x,y').$$

Indeed, assume $\varphi(x,y) \land \psi(x,y) \land x < y'$. Then we have

$$\begin{aligned} x &\leqslant y, && [\text{by } \psi(x,y)] \\ x = y &\lor x < y, && \\ x' = y' &\lor x' \leqslant y, && [\text{by } \varphi(x,y)] \\ x' &\leqslant y'. && \end{aligned}$$

This gives us (§). We shall use this to establish by induction

$$(\P) \qquad \forall y\ \forall x\ \bigl(\varphi(x,y) \land \psi(x,y)\bigr).$$

As the base of the induction, first we prove

$$(\|) \qquad \forall x\ \bigl(\varphi(x,1) \land \psi(x,1)\bigr).$$

Vacuously $\forall x\ \varphi(x,1)$, so by (‡) we have $\psi(1,x)$ and in particular $\psi(1,1)$. Using $\psi(1,x)$ and putting $(1,x)$ for (x,y) in (§) gives us

$$\begin{aligned} \varphi(1,x) &\Rightarrow \varphi(1,x'), \\ \psi(x,1) &\Rightarrow \psi(x',1). \end{aligned}$$

By induction, we have $\forall x\ \psi(x,1)$ and hence (‖). Now suppose we have for some y

$$(**) \qquad \forall x\ \bigl(\varphi(x,y) \land \psi(x,y)\bigr).$$

By (§) we get $\forall x\ \varphi(x,y')$. Finally we establish $\forall x\ \psi(x,y')$ by induction: From (‖) we have $\varphi(y',1)$, hence $\psi(1,y')$. Suppose $\psi(x,y')$. Then $\varphi(y',x)$. But from $\varphi(x,y')$ we have $\psi(y',x)$. Hence by (§) we have $\varphi(y',x')$, that is, $\psi(x',y')$. Thus, assuming (**), we have

$$\forall x\ \bigl(\varphi(x,y') \land \psi(x,y')\bigr).$$

By induction then, we have (¶). From this we extract $\forall y\ \forall x\ \psi(x,y)$, as desired. □

In Theorem 2, the condition (†) is not redundant. It is false that a structure that admits induction must have a linear ordering so that (†) is satisfied. It is true that all counterexamples to this claim are finite.[6] It may seem that there is no practical need to use induction in a finite structure, since the members can be checked individually for their satisfaction of some property. However, this checking may need to be done for every member of every set in an infinite family. Such is the case for Fermat's Theorem, as I have discussed elsewhere:

> Indeed, in the *Disquisitiones Arithmeticae* of 1801 [15, ¶50], which is apparently the origin of our notion of modular arithmetic, Gauss reports that Euler's first proof of Fermat's Theorem was as follows. Let p be a prime modulus. Trivially $1^p \equiv 1$ (with respect to p or indeed any modulus). If $a^p \equiv a$ (*modulo* p) for some a, then, since $(a+1)^p \equiv a^p + 1$, we conclude $(a+1)^p \equiv a+1$. This can be understood as a perfectly valid proof by induction in the ring with p elements that we denote by $\mathbb{Z}/p\mathbb{Z}$: we have then proved $a^p = a$ for all a in this ring. [33]

In analysis one learns some form of the following, possibly associated with the names of Heine and Borel.

Theorem 3. *The following are equivalent conditions on an interval I of $\mathbb{R}$:*
1. *I is closed and bounded.*
2. *All continuous functions from I to $\mathbb{R}$ are uniformly continuous.*

Such a theorem is pedagogically useful, both for clarifying the order of quantification in the definitions of continuity and uniform continuity, and for highlighting (or at least setting the stage for) the notion of compactness. A theorem about the equivalence of induction and well ordering serves no such useful purpose. If it is loaded up with enough conditions so that it is actually correct, as in Theorems 1 and 2 above, then there is only one structure (up to isomorphism) that meets the equivalent conditions, and this is just the usual structure of the natural numbers. If however the extra conditions are left out, as being a distraction to the immature

[6] Henkin investigates them in [20].

student, then that student may later be insensitive to the properties of structures like $\omega + \omega$ or $\mathbb{Z}/p\mathbb{Z}$. Thus the assertion that induction and well ordering are equivalent is nonrigorous in the worst sense: Not only does its proof require hidden assumptions, but the hiding of those assumptions can lead to real mathematical ignorance.

3 A standard of rigor

The highest standard of rigor might be the formal proof, verifiable by computer. But this is not a standard that most mathematicians aspire to. Normally one tries to write proofs that can be checked and *appreciated* by other human beings. In this case, ancient Greek mathematicians such as Euclid and Archimedes set an unsurpassed standard.

How can I say this? Two sections of Morris Kline's *Mathematical Thought from Ancient to Modern Times* [25] are called "The Merits and Defects of the *Elements*" (ch. 4, §10, p. 86) and "The Defects in Euclid" (ch. 42, §1, p. 1005). One of the supposed defects is, "he uses dozens of assumptions that he never states and undoubtedly did not recognize." I have made this criticism of modern textbook writers in the previous chapters, and I shall do so again in later chapters. However, it is not really a criticism unless the critic can show that bad effects follow from ignorance of the unrecognized assumptions. I shall address these assumptions in Euclid a bit later in this chapter.

Meanwhile, I think the most serious defect mentioned by Kline is the vagueness and pointlessness of certain definitions in Euclid. However, some if not all of the worst offenders were probably added to the *Elements* after Euclid was through with it.[1] Euclid apparently did not need these definitions. In any case, I am not aware that poor definitions make any proofs in Euclid confusing.

Used to prove the Side-Angle-Side Theorem (Proposition I.4) for triangle congruence, Euclid's method of superposition is considered a defect. However, if you do not want to use this method, then you can just make the theorem an axiom, as Hilbert does in *The Foundations of Geometry* [22]. (Hilbert's axioms are spelled out in Chapter 9 below.) I myself do not object to Euclid's proof by superposition. If two line segments are given as equal, what else can

[1] See in particular Russo's 'First Few Definitions in the *Elements*' [37, 10.15].

this mean but that one of them can be superimposed on the other? Otherwise equality would seem to be a meaningless notion. Likewise for angles. Euclid assumes that two line segments or two angles in a diagram *can* be given as equal. Hilbert assumes not only this, but something stronger: a given line segment can be *copied* to any other location, and likewise for an angle. Euclid *proves* these possibilities, as his Propositions I.3 and 23.

In his first proposition of all, where he constructs an equilateral triangle on a given line segment, Euclid uses two circles, each centered at one endpoint of the segment and passing through the other endpoint as in Figure 3.1. The two circles intersect at a point that

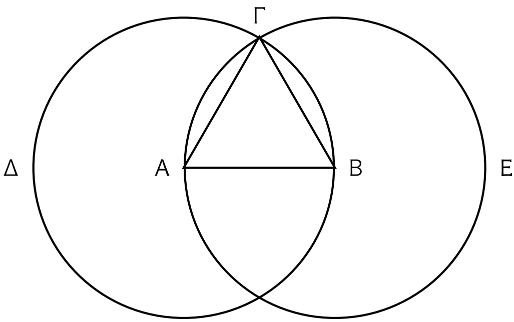


Figure 3.1: Euclid's Proposition I.1

is the apex of the desired triangle. But why should the circles intersect? It is considered a defect that Euclid does not answer this question. Hilbert avoids this question by not mentioning circles in his axioms.

Hilbert's axioms can be used to show that desired points on circles do exist. It is not a defect of rigor that Euclid does things differently. The original meaning of *geometry* in Greek is surveying. Herodotus [21, II.109] traces the subject to ancient Egypt, where the amount of land lost to the annual flooding of the Nile had to be measured. The last two propositions of Book I of Euclid are the Pythagorean

Theorem and its converse; but perhaps the climax of Book I comes two propositions earlier, with number 45. Here it is shown that, for a plot of land having any number of straight sides, an equal rectangular plot of land with one given side can be found. The whole point of Book I is to work out rigorously what can be done with tools such as a surveyor or perhaps a carpenter might have:

(1) a tool for drawing and extending straight lines;
(2) a tool for marking out the points that are equidistant from a fixed point; and
(3) a set square, not for *drawing* right angles, but for justifying the postulate that all right angles are equal to one another.

In the 19th century, it is shown that the same work can be accomplished with even less. This possibly reveals a defect of Euclid, but I do not think it is a defect of rigor.

There *is* however a danger in reading Euclid today. The danger lies in a hidden assumption; but it is an assumption that *we* make, not Euclid. We assume that, with his postulates, he is doing the same sort of thing that Hilbert is doing with *his* axioms. He is not. Hilbert has to deal with the possibility of non-Euclidean geometry. Hilbert can contemplate models that satisfy some of his axioms, but violate others. For Euclid, there is just one model: this world.

If mathematicians never encountered structures, other than the natural numbers, that were well ordered or admitted induction, then there might be nothing wrong with saying that induction and well ordering were equivalent. But then again, even to speak of equivalence is to suggest the possibility of *different* structures that satisfy the conditions in question. This is not a possibility that Euclid has to consider.

If Euclid is not doing what modern mathematicians are doing, what is the point of reading him? I respond that he is obviously doing *something* that we can recognize as mathematics. If he is just studying the world, so are mathematicians today; it is just a world that we have made more complicated. I suggested in the beginning that students are supposed to come to an analytic geometry class with some notion of synthetic geometry. As I observe below in Chapter 5 for example, students are *also* supposed to have the notion that every line segment has a length, which is a so-called real number. This is a notion that has been added to the world.

Nonetheless, the roots of this notion can be found in the *Elements,* in the theory of ratio and proportion, beginning in Book V.[2] According to this theory, magnitudes A, B, C, and D are **in proportion,** so that the **ratio** of A to B is the same as the ratio of C to D, if for all natural numbers k and m,

$$kA > mB \Leftrightarrow kC > mD.$$

In this case we may write

$$A : B :: C : D,$$

though Euclid uses no such notation. What is expressed by this notation is not the equality, but the *identity,* of two ratios. Equality is a possible property of two nonidentical magnitudes. Magnitudes are geometric things, ratios are not. Euclid never draws a ratio or assigns a letter to it.[3]

---

[2] According to a scholium to Book V, "Some say that the book is the discovery of Eudoxus, the pupil of Plato" [41, p. 409]. The so-called Euclidean Algorithm as used in Proposition X.2 of the *Elements* may be a remnant of another theory of proportion. Apparently this possibility was first recognized in 1933 [41, pp. 508–9]. The idea is developed in [14].

[3] I am aware of one possible counterexample to this claim. The last proposition—number 39—in Book VII is *to find the number that is the least of those that will have given parts.* The meaning of this is revealed in the proof, which begins: "Let the given parts be A, B, and $\mathsf{\Gamma}$. Then it is required to find the number that is the least of those that will have the parts A, B, and $\mathsf{\Gamma}$. So let $\mathsf{\Delta}$, E, and Z be numbers homonymous with the parts A, B, and $\mathsf{\Gamma}$, and let the least number H measured by $\mathsf{\Delta}$, E, and Z be taken." Thus H is the least common multiple of $\mathsf{\Delta}$, E, and Z, which can be found by Proposition VII.36. Also, if for example $\mathsf{\Delta}$ is the number n, then A is an nth, considered abstractly: it is not given as an nth part of anything in particular. Then A might be considered as the ratio of I to n. Possibly VII.39 was added later to Euclid's original text, although Heath's note [11, p. 344] suggests no such possibility. If indeed VII.39 is a later addition, then so, probably, are the two previous propositions, on which it relies: they are that if $n \mid r$, then r has an nth part, and conversely. But Fowler mentions Propositions 37 and 38, seemingly being as typical or as especially illustrative examples of propositions from Book VII [14, p. 359].

In any case, in the definition, it is assumed that A and B **have a ratio** in the first place, in the sense that some multiple of either of them exceeds the other; and likewise for C and D. In this case, the pair

$$\left(\left\{\frac{m}{k} : kA > mB\right\}, \left\{\frac{m}{k} : kA \leqslant mB\right\}\right)$$

is a *cut* (of positive rational numbers) in the sense of Dedekind [7, p. 13]. Dedekind traces his definition of irrational numbers to the conviction that

> an irrational number is defined by the specification of all rational numbers that are less and all those that are greater than the number to be defined... if... one regards the irrational number as the ratio of two measurable quantities then is this manner of determining it already set forth in the clearest possible way in the celebrated definition which Euclid gives of the equality of two ratios. [7, pp. 40–1]

In saying this, Dedekind intends to *distinguish* his account of the completeness or continuity of the real number line from some other accounts. Dedekind does not literally define an irrational number as a ratio of two "measurable quantities": the definition of cuts such as the one above does not require the use of magnitudes such as A and B. Dedekind observes moreover that Euclid's geometrical constructions do not require continuity of lines. "If any one should say" writes Dedekind

> that we cannot conceive of space as anything else than continuous, I should venture to doubt it and to call attention to the fact that a far advanced, refined scientific training is demanded in order to perceive clearly the essence of continuity and to comprehend that besides rational quantitative relations, also irrational, and besides algebraic, also transcendental quantitative relations are conceivable. [7, pp. 38]

Modern geometry textbooks (as in Chapter 5 below) assume continuity in this sense, but without providing the "refined scientific training" required to understand what it means. Euclid does provide something of this training, starting in Book V of the *Elements*; before this, he makes no use of continuity in Dedekind's sense.

In the Muslim and Christian worlds, Euclid has educated mathematicians for centuries. He shows the world what it means to prove things. One need not read *all* of the *Elements* today. But Book I lays out the basics of geometry in a beautiful way. If you want students to learn what a proof is, I think you can do no better then tell them, "A proof is something like what you see in Book I of the *Elements*."

I have heard of textbook writers who, informed of errors, decide to leave them in their books anyway, to keep the readers attentive. The perceived flaws in Euclid can be considered this way. The *Elements* must not be treated as a holy book. If it causes the student to think how things might be done better, this is good.

The *Elements* is not a holy book; it is one of the supreme achievements of the human intellect. It is worth reading for this reason, just as, say, Homer's *Iliad* is worth reading.

The rigor of Euclid's *Elements* is astonishing. Students in school today learn formulas, like $A = \pi r^2$ for the area of a circle. This formula encodes the following.

Theorem 4 (Proposition XII.2 of Euclid)**.** *Circles are to one another as the squares on the diameters.*

One might take this to be an obvious corollary of:

Theorem 5 (Proposition XII.1 of Euclid)**.** *Similar polygons inscribed in circles are to one another as the squares on the diameters.*

And yet Euclid gives an elaborate proof of XII.2 by what is today called the Method of Exhaustion:

Euclid's proof of Theorem 4, in modern notation. Suppose a circle C_1 with diameter d_1 is to a circle C_2 with diameter d_2 in a *lesser* ratio than ${d_1}^2$ is to ${d_2}^2$. Then ${d_1}^2$ is to ${d_2}^2$ as C_1 is to some fourth proportional R that is *smaller* than C_2. More symbolically,

$$C_1 : C_2 < {d_1}^2 : {d_2}^2,$$
$${d_1}^2 : {d_2}^2 : : C_1 : R,$$
$$R < C_2.$$

By inscribing in C_2 a square, then an octagon, then a 16-gon, and so forth, eventually (by Euclid's Proposition X.1) we obtain a 2^n-gon that is greater than R. The 2^n-gon inscribed in C_1 has (by Theorem 5, that is, Euclid's XII.1) the same ratio to the one inscribed in C_2 as ${d_1}^2$ has to ${d_2}^2$. Then

$${d_1}^2 : {d_2}^2 < C_1 : R,$$

which contradicts the proportion above. □

Such is Euclid's proof, in modern symbolism. Euclid himself does not refer to a 2^n-gon as such. His diagram must fix a value for 2^n, and the value fixed is 8. I do not know if anybody would consider this a lack of rigor, as if rigor is achieved by symbolism. I wonder how often modern symbolism is used to give only the *appearance* of rigor. (See Chapter 8 below.)

A more serious problem with Euclid's proof is the assumption of the existence of the fourth proportional R. Kline does not mention this as a defect in the sections of his book cited above; he does mention the assumption elsewhere (on page his 84), but not critically. Heath mentions the assumption in his own notes [11, v. 2, p. 375], though he does not supply the following way to avoid the assumption.

Second proof of Theorem 4. We assume instead that if two unequal magnitudes have a ratio in the sense of Book V of the *Elements,* then their *difference* has a ratio with either one of them.[4] In the notation above, since $C_1 : C_2 < {d_1}^2 : {d_2}^2$, there are some natural numbers m and k such that

$$mC_1 < kC_2, \qquad m{d_1}^2 \geqslant k{d_2}^2.$$

Let r be a natural number such that $r(kC_2 - mC_1) > C_2$. Then

$$rmC_1 < (rk-1)C_2, \qquad rm{d_1}^2 \geqslant rk{d_2}^2.$$

[4] This is the postulate of Archimedes: "among unequal [magnitudes], the greater exceeds the smaller by such a [difference] that is capable, added itself to itself, of exceeding everything set forth (of those which are in a ratio to one another" [5, p. 36].

Assuming $2^{n-1} \geqslant rk$, let P_1 be the 2^n-gon inscribed in C_1, and P_2 in C_2. Then

$$C_2 - P_2 < \frac{1}{2^{n-1}} C_2 \leqslant \frac{1}{rk} C_2,$$
$$rmP_1 < rmC_1 < (rk-1)C_2 < rkP_2.$$

But also $P_1 : P_2 :: {d_1}^2 : {d_2}^2$, so that $rmP_1 \geqslant rkP_2$, which is absurd. $\square$

Does this second proof of Theorem 4 supply a defect in Euclid's proof? In his note, Heath quotes Simson to the effect that assuming the mere existence of a fourth proportional does no harm, even if the fourth proportional cannot be constructed. I see no reason why Euclid could not have been aware of the possibility of avoiding this assumption, although he decided not to bother his readers with the details.

4 Why rigor

As used by Euclid and Archimedes, the Method of Exhaustion serves no practical purpose. Archimedes has an *intuitive* method [19] for finding equations of areas and volumes. He uses this method to discover that

(1) a section of a parabola is a third again as large as the triangle with the same base and height;
(2) if a cylinder is inscribed in a prism with square base, then the part of the cylinder cut off by a plane through a side of the top of the prism and the center of the base of the cylinder is a sixth of the prism; and
(3) the intersection of two cylinders is two thirds of the cube in which this intersection is inscribed.

However, Archimedes does not believe that his method provides a rigorous proof of his equations. He supplies proofs *after* the equations themselves are discovered. Why does he do this? After all, he believes Democritus should be credited for discovering that the pyramid is the third part of the prism with the same base and height, even though it was Eudoxus who later actually gave a proof. (The theorem is a corollary to Proposition XII.7 of Euclid's *Elements.*)

Although Heath translated Euclid faithfully into English, apparently he thought the rigor of Archimedes was too much for modern mathematicians to handle; so he paraphrased Archimedes with modern symbolism [17]. This symbolism is a way to avoid keeping too many ideas in one's head at once. When one wants to use a theorem for some practical purpose, then this labor-saving feature of symbolism is perhaps desirable. But if the whole point of a theorem is to see and *appreciate* something, then perhaps symbolism gets in the way of this.

I do not think Archimedes really explains his compulsion for mathematical rigor. Being the originator of the "merciless telegram style" that Landau [26, p. xi] for example writes in,[1] Euclid does not explain anything at all in the *Elements*; he just does the mathematics. I suppose the rigor of this mathematics, at least regarding proportions, is to be explained as a remnant of the discovery of incommensurable magnitudes. This discovery necessitates such a theory of proportion as is attributed to Eudoxus of Cnidus and is presented in Book V of the *Elements.* In any case, given a theory of proportion, if Euclid is going to *assert* Proposition XII.2, he is duty-bound to *prove* it in accordance with the theory at hand.

Modern mathematicians are likewise duty-bound to respect current standards of rigor. As suggested above in Chapter 2, this does not mean that a textbook has to prove everything from first principles; but at least some idea ought to be given of what those first principles are. This standard is set by Euclid and respected by Archimedes. It is not so much respected by modern textbooks of analytic geometry. In Chapters 5 and 8, I look at a couple of examples of these.

It may be said that the purpose of an analytic geometry text is to teach the student how to *do* certain things: how to solve certain problems, as detailed in the first quotation in Chapter 5 below. The purpose is not to teach proof. However, as suggested in Chapter 3 previous, Book I of the *Elements* is also concerned with doing things, with the help of such tools as a surveyor or carpenter might use. What makes the *Elements* mathematics is that it *justifies* the methods it gives for doing things. It provides proofs.

I suppose that, when a student of mathematics is given a problem, even a numerical problem, she or he is expected to be able to come up with a *solution,* and not just an *answer.* A solution is a proof that the answer is correct. It tells *why* the answer is correct. Thus it gives the reader the means to solve other problems. An analytic geometry text ought to prove that its methods are correct. At least it ought to give some indication of how the proofs might be supplied.

[1] Fowler [14, p. 386, n. 30] refers to Landau as the "premier exponent" of "a more recent German style of setting out mathematics, generally called 'Satz-Beweis' style, that has some affinities with *protasis*-style", that is, Euclid's style.

5 A book from the 1940s

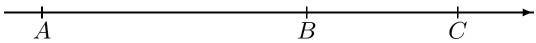
According to the Preface of Nelson, Folley, and Borgman's 1949 volume *Analytic Geometry* [31, pp. iii–iv]:[1]

> This text has been prepared for use in an undergraduate course in analytic geometry which is planned as preparation for the calculus rather than as a study of geometry. In order that it may be of maximum value to the future student of the calculus, the basic sciences, and engineering, considerable attention is given to two important problems of analytic geometry. They are (a) given the equation of a locus, to draw the curve, or describe it geometrically; (b) given the geometric description of a locus, to find its equation, that is, to translate a verbal description of a locus into a mathematical equation...
>
> Inasmuch as the student's ability to use analytic geometry as a tool depends largely on his understanding of the coordinate system, particular attention has been given to producing as thorough a grasp as possible. He must appreciate, for example, that the point (a, b) is not necessarily located in the first quadrant, and that the equation of a curve may be made to take a simple form if the coordinate axes are placed with forethought...

By referring to judicious placement of axes, the authors reveal their working hypothesis that there is already a geometric plane, before any coordinatization. It is not clear what students are expected to know about this plane. The book proper begins on page 3:

> **1. Directed Line Segments.** If A, B, and C (Fig. 1) are three points which are taken in that order on an infinite straight line,

[1] I have this book only because my mother used it in college. When I was twelve I used this book in order to plot graphs of conic sections. I also used the tables at the back to plot graphs of the trigonometric and logarithmic functions. I don't remember whether I tried to read the book from the beginning; if I did, I must not have got very far.

> then in conformity with the principles of plane geometry we may write
>
> $$AB + BC = AC. \tag{1}$$
>
> For the purposes of analytic geometry it is convenient to have equation (1) valid regardless of the order of the points A, B, and C on the infinite line. The conventional way of accomplishing this
>
> 
>
> FIG. 1
>
> is to select a positive direction on the line and then define the symbol AB to mean the number of linear units between A and B, or the negative of that number, according as we associate with the segment AB the positive or the negative direction. With this understanding the segment AB is called a *directed line segment.* In any given problem such a segment possesses an intrinsic sign decided in advance through the arbitrary selection of a positive direction for the infinite line of which the segment is a part...[2]

Thus it is assumed that the student knows what a "number of linear units" means. I suppose the student has been trained in high school to believe that (1) every line segment has a length and (2) this length is a number of some unit. But probably the student has no idea of how *numbers* in the original sense—natural numbers—can be used to create all of the numbers that might be needed to designate geometrical lengths.

Instead of making the unexamined assumption that every line segment has a numerical length, we could take *congruence* as the fundamental notion, and without defining length itself, we could say that congruent line segments have the *same length.* One who knew about equivalence relations could then define a length as a congruence class of segments; but this need not be made explicit.

Alternatively, and more in line with the approach of Nelson *& al.*, we can fix a unit line segment in the manner of Descartes in the

[2] This quotation has almost exactly the same visual appearance as in the original text. In particular, the lines and the placement of the figure are the same.

Geometry [8]. Then, by using the definition of proportion found in Book V of Euclid's *Elements* and discussed in Chapter 3 above, we can define the length of an arbitrary line segment as the ratio of this segment to the unit segment. This gives us lengths rigorously as real numbers, if we use Dedekind's definition of the latter. As Dedekind observed though, and as we repeated, there is no need to assume that *every* positive real number is the length of some segment.

Perhaps these details need not be rehearsed with the student. But neither is it necessary to introduce lengths at all in order to justify the equation $AB + BC = AC$ in the text. It need only be said that an expression like AB no longer represents merely a line segment, but a *directed* line segment. Then BA is the *negative* of AB, and we can write

$$(2) \qquad BA = -AB, \qquad AB = -BA, \qquad AB + BA = 0,$$

as indeed Nelson *& al.* do later in their §1 (on their page 4). Thus directed segments are understood to compose an abelian group. Indeed, although not a word is said about the commutativity or associativity of addition of segments, these properties might be understood to follow from the "principles of plane geometry" mentioned in the text.

If a positive direction is fixed for the straight line containing A and B, then AB itself is understood as positive, if B is further than A in the positive direction; otherwise AB is negative. Thus the abelian group of directed segments becomes an ordered group. I quote the next section of the text in its entirety.

> **2. Length, Distance.** The *length* of a directed line segment is the number of linear units which it contains. The symbol $|AB|$ will be used to designate the length of the segment AB, or the *distance between* the points A and B.
>
> Occasionally the symbol AB will be used to represent the line segment as a geometric entity, but if a numerical measure is implied then it stands for the directed segment AB or the *directed distance from A to B.*
>
> Two directed segments of the same line, or of parallel lines, are *equal* if they have equal lengths and the same intrinsic signs.

Again we see the unexamined assumption that the reader knows what a "number of linear units" means, when a length as such can be defined simply as a congruence class. Congruence of *directed* segments of a given straight line can be understood as a congruence of segments that is established by translation only, without reflection. The text defines this kind of congruence as equality.

An expression like AB in the text can now have any of three meanings. It can mean

(1) the segment bounded by A and B,
(2) the directed segment from A to B, or
(3) the directed distance from A to B, which we might identify with the class of directed segments that are equal to the directed segment from A to B.

In fact there will be yet another possible meaning of the expression AB, a meaning that will be used without comment in a quotation given below: AB can mean

(4) the infinite straight line containing A and B.

The second and third meanings of expressions like AB are used in exercises on page 7 of the book:

> In Problems 5–8, the consecutive points A, B, C, D, E, F, G are spaced one inch apart on an infinite line which is positively directed from A to B.
>
> **5.** Verify that $BD + GA = BA + GD$.
> **6.** Verify that $DB + GA = DA + GB$.
> **7.** Verify that $BG + FC = BC + FG = 2DE$.
> **8.** Verify that $\frac{1}{2}(EA + EG) = ED$.

Of the five equations here, the first three can be understood as equations of directed segments; the remaining two are equations of directed distances. The preamble to the problems here refers not just to abstract units, but to *inches,* which have no mathematical definition. Every mathematical statement about inches is just as true if inches are replaced with miles. This goes unmentioned.

In another problem, the authors display their assumption that numbers of units can be irrational and even transcendental:

> **11.** On a coordinate axis where the unit of measure is one

inch, plot the points whose coordinates are 2, $-\frac{5}{3}$, π, $3-\sqrt{5}$, and $\sqrt[3]{-16}$, respectively.

There is no properly *geometric* reason to introduce transcendental or even nonquadratic lengths. However, as noted in the quotation from the preface, the book is for students of the calculus. The book itself has a chapter called "Graphs of Single-Valued Transcendental Functions". Meanwhile, after a coordinatization of a straight line is considered in §3 of Chapter 1, the plane is considered:

4. Rectangular Coordinates. The coordinate system of the preceding article may be generalized so as to enable us to de-

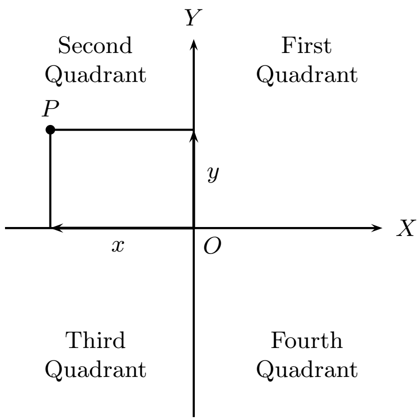


FIG. 5

scribe the location of a point in a the plane. Through any point O (Fig. 5) select two mutually perpendicular directed infinite straight lines OX and OY, thus dividing the plane into four parts called *quadrants,* which are numbered as shown in the figure. The point O is the *origin* and the directed lines are called the *x-axis* and the *y-axis,* respectively. A unit of measure is selected for each axis. Unless the contrary is stated, the units selected will be the same for both axes.

The directed distance from OY to any point P in the plane is the *x-coordinate,* or *abscissa,* of P; the directed distance from

> OX to the point is its *y-coordinate,* or *ordinate.* Together, the abscissa and ordinate of a point are called its *rectangular coordinates.* When a letter is necessary to represent the abscissa, x is most frequently used; y is used to represent the ordinate...
>
> ...Consequently, we may represent a point by its coordinates placed in parentheses (the abscissa always first), and refer to this symbol as the point itself. For example, we may refer to the point P_1 of [the omitted figure] as the point $(3, 5)$. Sometimes it is convenient to use both designations; we then write $P_1(3, 5)$... When a coordinate of a point is an irrational number, a decimal approximation is used in plotting the point...

Here OX and OY are not segments, but infinite straight lines. Nelson *& al.* evidently do not want to give a name such as $\mathbb{R}$ to the set of all numbers under consideration. Hence they cannot say that they identify the geometrical plane with $\mathbb{R} \times \mathbb{R}$; they can say only that they identify individual points with pairs of numbers. This is fine, except that it leaves unexamined the assumption that lengths are numbers of units.

How many generations of students have had to learn the words *abscissa* and *ordinate* without being given their etymological meanings?[3] Nelson *& al.* do not discuss them, even in their chapter on conic sections, although the terms are the Latin translations of Greek words used by Apollonius of Perga in the *Conics* [3]. I consider their original meaning in Chapter 6 below.

Meanwhile, I just have to wonder whether an analytic geometry textbook cannot be more enticing than that of Nelson *& al.* If the purpose of the subject is to solve problems, why not present some of the actual problems that the subject was invented to solve? A possible example is the duplication of the cube, discussed at the end of Chapter 7.

[3] One complaint I have about my own education is having had to learn technical terms without their etymology. In a literature class in high school, how much easier it would have been to learn what *zeugma* was, if only it had been pointed out to us that the Greek word was cognate with the Latin-derived *join* and the Anglo-Saxon *yoke.*

6 Abscissas and ordinates

In the first of the eight books[1] of the *Conics* [3], Apollonius derives properties of the conic sections that can be used to write their equations in rectangular or oblique coordinates. I review these properties here, because (1) they have intrinsic interest, (2) they are the reason why Apollonius gave to the three conic sections the names that they now have, and (3) the vocabulary of Apollonius is a source for many of our technical terms, including *abscissa* and *ordinate.*

Apollonius did not invent any of his terms: these were just ordinary words, used in a certain way. When we carry those words—or their Latin versions—over into our own language, we create some distance between ourselves and mathematics. When I first learned that a conic section had a *latus rectum,* I understood that there was a whole theory of conic sections that was not being revealed, although its existence was hinted at by this peculiar Latin term. Had the *latus rectum* been called an upright side as in Apollonius, it would have been easier to ask "What is an upright side?" In turn, textbook writers might have felt more obliged to explain what it was. In any case, I am going to give an explanation below.

English does borrow foreign words freely: this is a characteristic of the language. A large lexicon is not a bad thing. A choice from among two or more synonyms can help establish the register of a piece of speech.[2] If distinctions between near-synonyms are carefully maintained, then subtlety of expression is possible. *Circle* and *cycle* are Latin and Greek words for the same thing, but the Greek word

[1] The first four books survive in Greek, the next three in Arabic translation; the last book is lost.

[2] In the 1980s, the *Washington Post* described a best-selling book called *Color Me Beautiful* as offering "the color-wheel approach to female pulchritude." The *New York Times* just said the book provided "beauty tips for women." The register of the *Post* was mocking; the *Times,* neutral.

is used more abstractly in English, and it would be bizarre to refer to a finite group of prime order as being "circular."

However, mathematics can be done in any language. Greek does mathematics without a specialized vocabulary. It is worthwhile to consider what this is like.

For Apollonius, a **cone** (ὁ κῶνος "pine-cone") is a solid figure determined by (1) a **base** (ἡ βάσις), which is a circle, and (2) a **vertex** (ἡ κορυφή "summit"), which is a point that is not in the plane of the base. The surface of the cone contains all of the straight lines drawn from the vertex to the circumference of the base. A **conic surface** (ἡ κωνικὴ ἐπιφάνεια[3]) consists of such straight lines, not bounded by the base or the vertex, but extended indefinitely in both directions.

The straight line drawn from the vertex of a cone to the center of the base is the **axis** (ὁ ἄξων "axle") of the cone. If the axis is perpendicular to the base, then the cone is **right** (ὀρθός); otherwise it is **scalene** (σκαληνός "uneven"). Apollonius considers both kinds of cones indifferently.

A plane containing the axis intersects the cone in a triangle. Suppose a cone with vertex A has axial triangle ABC. Then the base BC of this triangle is a diameter of the base of the cone. Let an arbitrary chord[4] DE of the base of the cone cut the base BC of the axial triangle at right angles at a point F, as in Figure 6.1. In the axial triangle, let the straight line FG be drawn from the base to the side AC. This straight line FG may, but need not, be parallel to the side BA. It is not at right angles to DE, unless the plane of the axial triangle is at right angles to the plane of the base of the cone. In any case, the two straight lines FG and DE, meeting at F, are not in a straight line with one another, and so they determine a plane. This plane cuts the surface of the cone in such a curve

[3]The word ἐπιφάνεια means originally "appearance" and is the source of the English "epiphany."

[4]Although it is the source of the English *cord* and *chord* [23], Apollonius does not use the word ἡ χορδή, although he proves in Proposition I.10 that the straight line joining any two points of a conic section *is* a chord, in the sense that it falls within the section. The Greek χορδή means gut, hence *anything* made with gut, be it a lyre-string or a sausage [27].

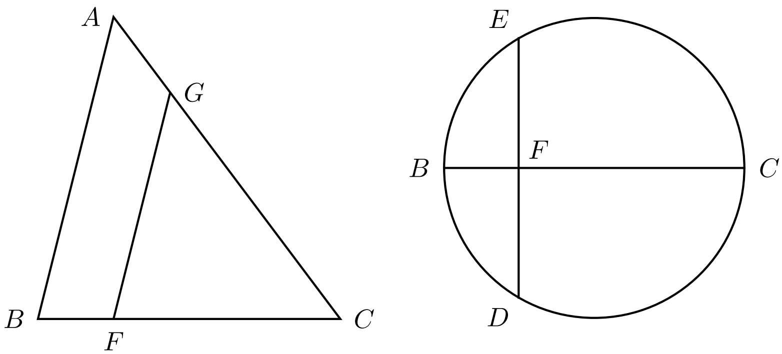


Figure 6.1: Axial triangle and base of a cone

DGE as is shown in Figure 6.2. Apollonius refers to such a curve

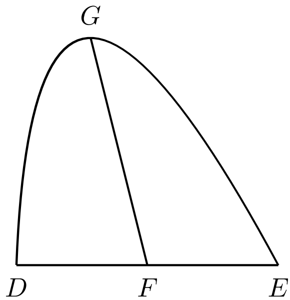


Figure 6.2: A conic section

first (in Proposition I.7) as a section (ἡ τομή) in the *surface* of the cone, and later (I.10) as a section of a cone. All of the chords of this section that are parallel to DE are bisected by the straight line GF. Therefore Apollonius calls this straight line a **diameter** (ἡ διάμετρος [γραμμή]) of the section.

The parallel chords bisected by the diameter are said to be drawn to the diameter **in an orderly way.** The Greek adverb here is τεταγμένως [2], from the verb τάσσω, which has meanings like “to draw up in order of battle" [27]. A Greek noun derived from this verb is τάξις, which is found in English technical terms like *taxonomy* and *syntax* [28]. The Latin adverb corresponding the Greek τεταγμένως is *ordinate* from the verb *ordino.* From the Greek expression for *the straight line drawn in an orderly way,* Apollonius will elide the middle part, leaving *the in-an-orderly-way.*[5] This term will refer to *half* of a chord bisected by a diameter. Similar elision in the Latin leaves us with the word **ordinate** for this half-chord [30]. Descartes refers to ordinates as *[lignes] qui s'appliquent par ordre [au] diametre* [8, p. 328].

The point G at which the diameter GF cuts the conic section DGE is called a **vertex** (κορυφής as before). The segment of the diameter between the vertex and an ordinate has come to be called in English an **abscissa;** but this just the Latin translation of Apollonius's Greek for being cut off (ἀπολαμβανομένη "taken"[6]).

Apollonius will show that every point of a conic section is the vertex for some unique diameter. If the ordinates corresponding to a particular diameter are at right angles to it, then the diameter will be an **axis** of the section. Meanwhile, in describing the relation between the ordinates and the abscissas of conic section, there are three cases to consider.

The parabola

Suppose the diameter of a conic section is parallel to a side of the corresponding axial triangle. For example, suppose in Figure 6.1 that FG is parallel to BA. The square on the ordinate DF is equal to the rectangle whose sides are BF and FC (by Euclid's Proposition III.35). More briefly, $DF^2 = BF \cdot FC$. But BF is independent of

[5]Heath [1, p. clxi] translates τεταγμένως as *ordinate-wise*; Taliaferro [3, p. 3], as *ordinatewise.* But this usage strikes me as anachronistic. The term *ordinatewise* seems to mean *in the manner of an ordinate*; but ordinates are just what we are trying to define when we translate τεταγμένως.

[6]I note the usage of the Greek participle in [2, I.11, p. 38]. Its general usage for what we translate as *abscissa* is confirmed in [27], although the general sense of the verb is not of cutting, but of taking.

the choice of the point D on the conic section. That is, for any such choice (aside from the vertex of the section), a plane containing the chosen point and parallel to the base of the cone cuts the cone in another circle, and the axial triangle cuts this circle along a diameter, and the plane of the section cuts this diameter at right angles into two pieces, one of which is equal to BF. The square on DF thus varies as FC, which varies as FG. That is, the square on an ordinate varies as the abscissa (I.20). Hence there is a straight line GH such that

$$DF^2 = FG \cdot GH,$$

and GH is independent of the choice of D.

This straight line GH can be conceived as being drawn at right angles to the plane of the conic section DGE. Apollonius calls GH the **upright side** (ὀρθία [πλευρά]), and Descartes accordingly calls it *le costé droit* [8, p. 329]. Apollonius calls the conic section itself a **parabola** (ἡ παραβολή), that is, an *application,* presumably because the rectangle bounded by the abscissa and the upright side is the result of *applying* (παραβάλλω) the square on the abscissa to the upright side. Such an application is made for example in Proposition I.44 of Euclid's *Elements,* where a parallelogram equal to a given triangle is applied to a given straight line. (This proposition is a lemma for Proposition 45, mentioned in Chapter 3 above as the climax of Book I of the *Elements.*)

The Latin term for upright side is *latus rectum.* This term is also used in English. In the *Oxford English Dictionary,* the earliest quotation illustrating the use of the term is from a mathematical dictionary published in 1702. Evidently the quotation refers to Apollonius and gives his meaning:

> App. Conic Sections 11 In a Parabola the Rectangle of the Diameter, and Latus Rectum, is equal to the rectangle of the Segments of the double Ordinate. [30]

I assume the "segments of the double ordinate" are the two halves of a chord, so that each of them is what we are calling an ordinate, and the rectangle contained by them is equal to the square on one of them.

The textbook by Nelson *& al.* considered in Chapter 5 above defines the parabola in terms of a *focus* and *directrix.* The possibility of defining all of the conic sections in this way is demonstrated by Pappus [32, p. 1012] and was probably known to Apollonius.[7] According to Nelson *& al.*,

> The chord of the parabola which contains the focus and is perpendicular to the axis is called the *latus rectum.* Its length is of value in estimating the amount of "spread" of the parabola.

The first sentence here defines the *latus rectum* so that it is four times the length of Apollonius's. The second sentence correctly describes the significance of the *latus rectum.* However, the juxtaposition of the two sentences may mislead somebody who knows just a little Latin. The Latin adjective *latus, -a, -um* does mean "broad, wide; spacious, extensive": it is the root of the English noun *latitude.* However, the Latin adjective *latus* is unrelated to the noun *latus, -eris* "side; flank" [29], which is found in English in the adjective *lateral*; and this noun is what is used in the phrase *latus rectum.*[8]

[7] Each conic section can be understood as the locus of a point whose distance from a given point has a given ratio to its distance from a given straight line. As Heath [1, pp. xxxvi–xl] explains, Pappus proves this theorem because Euclid did not supply a proof in his now-lost treatise on *surface loci.* Euclid must have omitted the proof because it was already well known; and Euclid predates Apollonius. Kline [25, p. 96] summarizes all of this by saying that the focus-directrix property "was known to Euclid and is stated and proved by Pappus." Later (on his page 128) Kline gives the precise reference to Pappus: it is Proposition 238 [in Hultsch's numbering] of Book VII. "As noted in the preceding chapter" he says, "Euclid probably knew it."

[8] In *latus rectum,* the adjective *rectus, -a, -um* "straight, upright" is given the neuter form, because the noun *latus* is neuter. The plural of *latus rectum* is *latera recta.* The neuter plural of the adjective *latus* would be *lata.* The dictionary writes the adjective as *lātus,* with a long "a"; but the "a" in the noun is unmarked and therefore short. As far as I can tell, the adjective is to be distinguished from another Latin adjective with the same spelling (and the same long "a"), but with the meaning of "carried, borne", used for the past participle of the verb *fero, ferre, tulī, lātum.* This past participle appears in English in words like *translate,* while *fer-* appears in *transfer.* The *American Heritage Dictionary* [28] traces *lātus* "broad" to an Indo-European root *stel-* and gives *latitude* and *dilate* as English derivatives; *lātus* "carried" comes from an Indo-European root *tel-* and is found in English

Denoting abscissa by x, and ordinate by y, and *latus rectum* by ℓ, we have for the parabola the modern equation

$$y^2 = \ell x. \qquad (*)$$

The hyperbola

The second possibility for a conic section is that the diameter meets the other side of the axial triangle when this side is extended beyond the vertex of the cone. In Figure 6.3, the diameter FG, crossing one

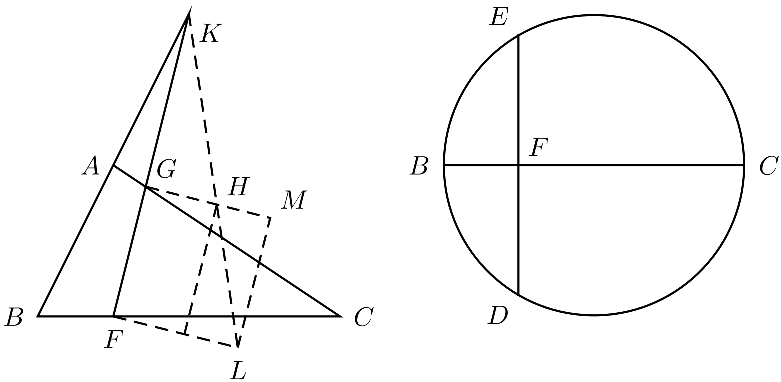


Figure 6.3: Axial triangle and base of a cone

side of the axial triangle ABC at G, crosses the other side, extended,

words like *translate* and *relate,* but also *dilatory.* Thus *dilatory* is not to be considered as a derivative of *dilate.* A French etymological dictionary [10] implicitly confirms this under the adjacent entries *dilater* and *dilatoire.* The older Skeat [39] does give *dilatory* as a derivative of *dilate.* However, under *latitude,* he traces *lātus* "broad" to the Old Latin *stlātus,* while under *tolerate* he traces *lātum* "borne" to *tlātum.* In his introduction, Skeat says he has collated his dictionary "with the *New English Dictionary* [as the *Oxford English Dictionary* was originally called] from A to H (excepting a small portion of G)." In fact the *OED* distinguishes *two* English verbs *dilate,* one for each of the Latin adjectives *lātus.* But the dictionary notes, "The sense 'prolong' comes so near 'enlarge', 'expand', or 'set forth at length'... that the two verbs were probably not thought of as distinct words."

at K. Again $DF^2 = BF \cdot FC$; but the latter product now varies as $KF \cdot FG$. The upright side GH can now be defined so that

$$BF \cdot FC : KF \cdot FG : : GH : GK.$$

We draw KH and extend to L so that FL is parallel to GH, and we extend GH to M so that LM is parallel to FG. Then

$$\begin{aligned} FL \cdot FG : KF \cdot FG &: : FL : KF \\ &: : GH : GK \\ &: : BF \cdot FC : KF \cdot FG, \end{aligned}$$

and so $FL \cdot FG = BF \cdot FC$. Thus

$$DF^2 = FG \cdot FL.$$

Apollonius calls the conic section here an **hyperbola** (ἡ ὑπερβολή), that is, an *exceeding,* because the square on the ordinate is equal to a rectangle whose one side is the abscissa, and whose other side is applied to the upright side; but this rectangle *exceeds* (ὑπερβάλλω) the rectangle contained by the abscissa and the upright side by another rectangle. This last rectangle is similar to the rectangle contained by the upright side and GK. Apollonius calls GK the **transverse side** (ἡ πλαγία πλευρά) of the hyperbola. Denoting its length by a, and the other segments as before, we have the modern equation

$$y^2 = \ell x + \frac{\ell}{a} x^2. \tag{†}$$

The ellipse

The last possibility is that the diameter meets the other side of the axial triangle when this side is extended below the base. All of the computations will be as for the hyperbola, except that now, if it is considered as a *directed* segment as in Chapter 5, the transverse side is negative, and so the modern equation is

$$y^2 = \ell x - \frac{\ell}{a} x^2. \tag{‡}$$

In this case Apollonius calls the conic section an **ellipse** (ἡ ἔλλειψις), that is, a *falling short,* because again the square on the ordinate is equal to a rectangle whose one side is the abscissa, and whose other side is applied to the upright side; but this rectangle now *falls short* (ἐλλείπω) of the rectangle contained by the abscissa and the upright side by another rectangle. Again this last rectangle is similar to the rectangle contained by the upright and transverse sides.

Thus the terms *abscissa* and *ordinate* are ultimately translations of Greek words that merely describe certain line segments that can be used to describe points on conic sections. For Apollonius, they are not required to be at right angles to one another.

Descartes generalizes the use of the terms slightly. In one example [8, p. 339], he considers a curve derived from a given conic section in such a way that, if a point of the conic section is given by an equation of the form

$$y^2 = \dots x \dots,$$

then a point on the new curve is given by

$$y^2 = \dots x' \dots,$$

where xx' is constant. But Descartes just describes the new curve in words:

> toutes les lignes droites appliquées par ordre a son diametre estant esgales a celles d'une section conique, les segmens de ce diametre, qui sont entre le sommet & ces lignes, ont mesme proportion a une certaine ligne donnée, que cete ligne donnée a aux segmens du diametre de la section conique, auquels les pareilles lignes sont appliquées par ordre.[9]

There is still no notion that an arbitrary point might have two coordinates, called abscissa and ordinate respectively.

[9] "All of the straight lines drawn in an orderly way to its diameter being equal to those of a conic section, the segments of this diameter that are between the vertex and these lines have the same ratio to a given line that this given line has to the segments of the diameter of the conic section to which the parallel lines are drawn in an orderly way."

7 The geometry of the conic sections

For an hyperbola or ellipse, the **center** (κέντρον) is the midpoint of the transverse side. In Book I of the *Conics,* Apollonius shows that the diameters of (1) an ellipse are the straight lines through its center, (2) an hyperbola are the straight lines through its center that actually cut the hyperbola,[1] (3) a parabola are the straight lines that are parallel to the axis. Moreover, with respect to a new diameter, the relation between ordinates and abscissas is as before, except that the upright and transverse sides may be different.

I do not know of an efficient way to prove these theorems by Cartesian, analytic methods. Descartes opens his *Geometry* by saying,

> All problems in geometry can easily be reduced to such terms that one need only know the lengths of certain straight lines in order to solve them.

However, Apollonius proves his theorems about diameters by means of *areas.* Areas can be reduced to products of straight lines, but the reduction in the present context seems not to be particularly easy. For example, to shift the diameter of a parabola, Apollonius will use the following.

Lemma 3 (Proposition I.42 of Apollonius). *In Figure 7.1, it is assumed that (1) the parabola GDB has diameter AB, (2) AG is tangent to the parabola at G, (3) GJ is an ordinate, and (4) $GJBH$ is a parallelogram. Moreover (5) the point D is chosen at random on the parabola, and (6) triangle EDZ is drawn similar to AGJ. It follows that*

the triangle EDZ is equal to the parallelogram HZ.

[1] If the hyperbola is considered together with its conjugate hyperbola, then all straight lines through the center are diameters, except the asymptotes.

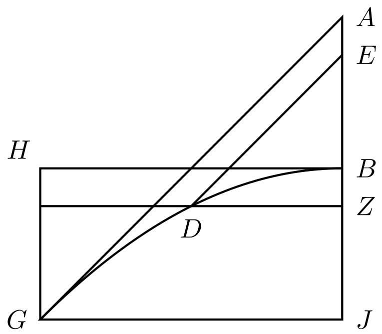


Figure 7.1: Proposition I.42 of Apollonius

Proof. The proof relies on knowing (from I.35) that $AB = BJ$. Therefore $AGJ = HJ$. Thus the claim follows when D is just the point G. In general we have

$$\begin{aligned} EDZ : HJ &:: EDZ : AGJ && \text{[Euclid V.7]} \\ &:: DZ^2 : GJ^2 && \text{[Euclid VI.19]} \\ &:: BZ : BJ && \text{[Apollonius I.20]} \\ &:: HZ : HJ, && \text{[Euclid VI.1]} \end{aligned}$$

and so $EDZ = HZ$ by Euclid V.8. The relative positions of D and G on the parabola are irrelevant to the argument. □

Then the diameter of a parabola can be shifted by the following.

Theorem 6 (Proposition I.49 of Apollonius)**.** *In Figure 7.2, it is assumed that (1) KDB is a parabola, (2) its diameter is MBG, (3) GD is tangent to the parabola, and (4) through D, parallel to BG, straight line ZDN is drawn. Moreover (5) the point K is chosen at random on the parabola, (6) through K, parallel to GD, the straight line KL is drawn, and (7) BR is drawn parallel to GD. It follows*

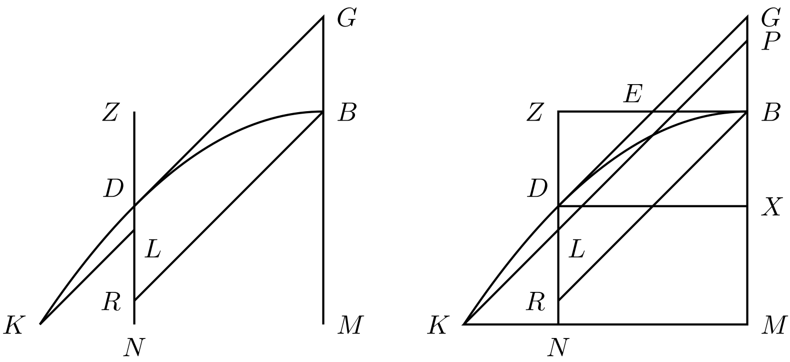


Figure 7.2: Proposition I.49 of Apollonius

that[2]

$$KL^2 : BR^2 : : DL : DR.$$

Proof. Let ordinate DX be drawn, and let BZ be drawn parallel to it. Then

$$\begin{aligned} GB &= BX && \text{[Apollonius I.35]} \\ &= ZD, && \text{[Euclid I.34]} \end{aligned}$$

and so (by Euclid I.26 & 29)

$$\triangle EGB = \triangle EZD.$$

Let ordinate KNM be drawn. Adding to either side of the last equation the pentagon $DEBMN$, we have the trapezoid $DGMN$ equal to the parallelogram ZM (that is, $ZBMN$).

Let KL be extended to P. By the lemma above, the parallelogram ZM is equal to the triangle KPM. Thus

$$DGMN = KPM.$$

[2]Apollonius also finds the upright side corresponding to the new diameter DN: it is H such that $ED : DZ : : H : 2GD$.

Subtracting the trapezoid $LPMN$ gives

$$KLN = LG.$$

We have also

$$BRZ = RG$$

(as by adding the trapezoid $DEBR$ to the equal triangles EZD and EGB). Therefore

$$\begin{aligned} KL^2 : BR^2 &:: KLN : BRZ \\ &:: LG : RG \\ &:: LD : RD. \end{aligned} \qquad \square$$

The proof given above works when K is to the left of D. The argument can be adapted to the other case. Then, as a corollary, we have that DN bisects all chords parallel to DG. In fact Apollonius proves this independently, in Proposition I.46.

Again, I do not see how the foregoing arguments can be improved by expressing all of the areas involved in terms of lengths. Rule Four in Descartes's *Rules for the Direction of the Mind* [9] is, "We need a method if we are to investigate the truth of things." Descartes elaborates:

> ...So useful is this method that without it the pursuit of learning would, I think, be more harmful than profitable. Hence I can readily believe that the great minds of the past were to some extent aware of it, guided to it even by nature alone...This is our experience in the simplest of sciences, arithmetic and geometry: we are well aware that the geometers of antiquity employed a sort of analysis which they went on to apply to the solution of every problem, though they begrudged revealing it to posterity. At the present time a sort of arithmetic called "algebra" is flourishing, and this is achieving for numbers what the ancients did for figures...But if one attends closely to my meaning, one will readily see that ordinary mathematics is far from my mind here, that it is quite another discipline I am expounding, and that these illustrations are more its outer garments than its inner parts...Indeed, one can even see some traces of this true mathematics, I think,

> in Pappus and Diophantus who, though not of that earliest antiquity, lived many centuries before our time. But I have come to think that these writers themselves, with a kind of pernicious cunning, later suppressed this mathematics as, notoriously, many inventors are known to have done where their own discoveries are concerned... In the present age some very gifted men have tried to revive this method, for the method seems to me to be none other than the art which goes by the outlandish name of "algebra"—or at least it would be if algebra were divested of the multiplicity of numbers and imprehensible figures which overwhelm it and instead possessed that abundance of clarity and simplicity which I believe true mathematics ought to have.

Descartes does not mention Apollonius among the ancient mathematicians, and I do not believe that in his *Geometry* he has managed to recover the method whereby Apollonius proves all of his theorems.

On the other hand, Descartes may have recovered *one* method used by ancient mathematicians, because perhaps some of these mathematicians *did* solve problems by considering equations of polynomial functions of lengths only. An example is Menaechmus, "a pupil of Eudoxus and a contemporary of Plato" [1, p. xix].

Apollonius did not discover the conic sections; Menaechmus is thought to have done this, if only because his is the oldest name associated with the conic sections. According to the commentary by Eutocius[3] on Archimedes, Menaechmus had two methods for finding two mean proportionals to two given straight lines; each of these methods uses conic sections. One of the methods is illustrated by Figure 7.3; apparently Menaechmus's own diagram was just like this [5, p. 288]. Given the lengths A and E, we want to find B and Γ so that

$$\mathsf{A} : \mathsf{B} :: \mathsf{B} : \Gamma :: \Gamma : \mathsf{E},$$

or equivalently

$$(*) \qquad \mathsf{B}^2 = \mathsf{A} \cdot \Gamma, \qquad\qquad \mathsf{B} \cdot \Gamma = \mathsf{A} \cdot \mathsf{E}.$$

[3]Eutocius flourished around 500 C.E., and his commentary was revised by Isadore of Miletus [18, p. 25], who along with Anthemius of Tralles was a master-builder of Justinian's Ayasofya [35].

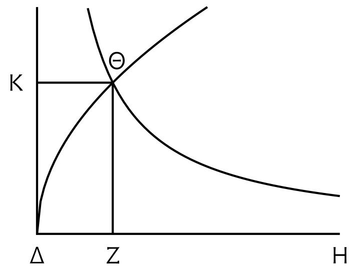


Figure 7.3: Menaechmus's finding of two mean proportionals

In the special case where A is twice E, we shall have that the cube with side Γ is double the cube with side E. In any case, it is sufficient if (1) B is an ordinate, and Γ the corresponding abscissa, of the parabola with upright side A whose axis is ΔH in the diagram, and (2) B and Γ are the coordinates of a point on the hyperbola whose asymptotes are ΔK and ΔH in the diagram and which also passes through the points with coordinates A and E. For then we shall have (∗) as desired. Thus, if Θ is the intersection of the parabola and hyperbola, we can let B be ZΘ and let Γ be ΔZ.

We have used the property proved by Apollonius in his Proposition II.12, that the rectangle bounded by the straight lines drawn from a point on an hyperbola to the asymptotes has constant area. Heath has an idea of how Menaechmus proved this [1, xxv–xxviii]. In any case, by the report of Eutocius, Menaechmus's other method of finding two mean proportionals was to use two parabolas with orthogonal axes.

I referred to B and Γ as coordinates, but this is an anachronism. According to one historian,

> Since this material has a strong resemblance to the use of coordinates, as illustrated above, it has sometimes been maintained that Menaechmus had analytic geometry. Such a judgment is warranted only in part, for certainly Menaechmus was unaware that any equation in two unknown quantities determines a curve. In fact, the general concept of an equation in unknown quantities

> was alien to Greek thought. It was shortcomings in algebraic notations that, more than anything else, operated against the Greek achievement of a full-fledged coordinate geometry. [6, pp. 104–5][4]

Boyer evidently considers analytic geometry as the study of the graphs of arbitrary equations; but this would seem to be within the purview of calculus rather than geometry. The book of Nelson *& al.* discussed in Chapter 5 does have chapters on graphs of single-valued algebraic functions, single-valued transcendental functions, and multiple-valued functions, as well as on parametric equations; but this fits the explicit purpose of the text as a preparation for calculus.

Did Descartes have a "full-fledged analytic geometry" in the sense of Boyer? In the *Geometry* [8, pp. 315–7], Descartes rejects the study of curves like the quadratrix, which today can be defined by the equation

$$\tan\left(\frac{\pi}{2}\cdot y\right) = \frac{y}{x},$$

or more elaborately by the pair of equations

$$\frac{\theta}{y} = \frac{\pi}{2}, \qquad \tan\theta = \frac{y}{x},$$

the variables being as in Figure 7.4. Descartes does not write down an equation for the quadratrix; but an equation is not needed for proving theorems about this curve. Pappus [41, pp. 336–47] defines the quadratrix as being traced in a square by the intersection of two straight lines, one horizontal and moving from the top edge BG to the bottom edge AD, the other swinging about the lower left corner A from the left edge AB to the bottom edge AD. Assuming there is

[4] In fact what Boyer refers to as "this material" is the properties of the conic sections given by equations ($*$), ($\dagger$) and ($\ddagger$) in the previous chapter. Boyer will presently give the method of cube-duplication using two parabolas, and then say, "It is probable that Menaechmus knew that the duplication could be achieved also by the use of a rectangular hyperbola and a parabola." It is not clear why he says "It is probable that," unless he questions the authority of Eutocius.

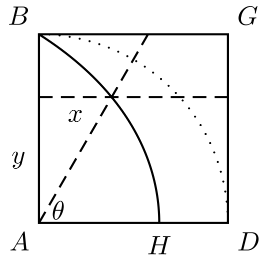


Figure 7.4: The quadratrix

a point H where the quadratrix meets the lower edge of the square, we have

$$BD : AB : : AB : BH,$$

where BD is the circular arc centered at A. Then a straight line equal to this arc can be found, and so the circle can be squared. This is why the curve is called the quadratrix (τετραγωνίζουσα). Pappus demonstrates this property, while pointing out that we have no way to construct the quadratrix without knowing where the point H is in the first place.[5] Today we have a notation for its position: if D is one unit away from A, then the length of AH is what we call $2/\pi$. However, this notation does not give us the location of H any better than Pappus's description of the quadratrix does. Today we can prove that π (and hence $2/\pi$) is transcendental; but this is not a topic of analytic geometry.[6]

Is "the general concept of an equation in unknown quantities" something that is "alien to Greek thought"? Perhaps it is alien to our own thought. According to Boyer, "any equation in two unknown quantities determines a curve." But this would seem to be an exaggeration, unless an arbitrary subset S of the plane $\mathbb{R} \times \mathbb{R}$ is to be considered a curve. For, if χ_S is the characteristic function of S,

[5] Pappus attributes this criticism to one Sporus, about whom we apparently have no source but Pappus himself [5, p. 285, n. 78].

[6] It is however a topic included in Spivak's *Calculus* [40, ch. 20].

then S is the solution-set of the equation

$$\chi_S(x, y) = 1.$$

Probably Boyer does not have in mind equations with parameters like S, but equations whose only parameters are real numbers, and in particular equations that are expressed by means of polynomial, trigonometric, logarithmic, and exponential functions.

If Menaechmus neglects to study all such functions, it is not for lack of adequate algebraic notation, but lack of interest. He solves the problem of finding two mean proportionals to two given line segments. If a numerical approximation is wanted, this can be found, as close as desired; therefore, by the continuity of the real line established by Dedekind, an exact solution exists. But Menaechmus wants a *geometric* solution, and he finds one, evidently by using the kind of mathematics that we refer to today as analytic geometry. Indeed, Heath suspects that Menaechmus first came up with the equations $(*)$ and *then* discovered that curves defined by these equations could be obtained as conic sections [1, p. xxi]. Figure 7.3 could appear at the beginning of any analytic geometry text, as an illustration of what the subject is about.

Pappus [41, pp. 346–53] reports three kinds of geometry problem: **plane,** as being solved by means straight lines and circles only, which lie in a plane; **solid,** as requiring also the use of conic sections, which in particular are sections of a solid figure; and **linear,** as involving more complicated *lines,* that is, curves, such as the quadratrix. Perhaps justly, Descartes criticizes this analysis as simplistic. He shows that curves given by polynomial equations have a heirarchy determined by the degrees of the polynomials. This hierarchy could have been meaningful for Pappus, since lower-degree curves can be used to construct higher-degree curves by methods more precise than the construction of the quadratrix.

One solid problem described by Pappus [41, pp. 486–9] is the four-line locus problem: find the locus of points such that the rectangle whose dimensions are the distances to two given straight lines bears a given ratio to the rectangle whose dimensions are the distances to two more given straight lines. According to Pappus, theorems of Apollonius were needed to solve this problem; but it is not clear whether Pappus thinks Apollonius actually did work out a full solution. By the last three propositions, namely 54–6, of Book III of the *Conics* of Apollonius, it is implied that the conic sections are three-line loci, that is, solutions to the four-line locus problem when two of the lines are identical. Taliaferro [3, pp. 267–75] works out the details and derives the theorem that the conic sections are four-line loci.

Descartes works out a full solution to the four-line locus problem. He also solves a particular *five*-line locus problem: the solution is a curve obtained as the intersection of a sliding parabola and a straight line through two points, one fixed, the other sliding along with the parabola.

Thus Descartes would seem to have made progress along an ancient line of research, rather than just heading off in a different direction. As Descartes observes, Pappus [42, pp. 600-3] could *formulate* the $2n$-line locus problem for arbitrary n. If $n > 3$, the ratio of the product of n segments with the product of n segments can be understood as the ratio compounded of the respective ratios of segment to segment. That is, given $2n$ segments A_1, …, A_n, B_1, …, B_n, we can understand the ratio of the product of the A_k to the product of the B_k as the ratio of A_1 to C_n, where

$$\begin{gathered} A_1 : C_1 : : A_1 : B_1, \\ C_1 : C_2 : : A_2 : B_2, \\ C_2 : C_3 : : A_3 : B_3, \\ \dots\dots\dots\dots\dots\dots, \\ C_{n-1} : C_n : : A_n : B_n. \end{gathered}$$

Descartes expresses the solution of the $2n$-line locus problem as an nth-degree polynomial equation in x and y, where y is the distance from the point to one of the given straight lines, and x is the distance from a given point on that line to the foot of the perpendicular from the point of the locus.

In fact Descartes does not use the perpendicular as such, but a straight line drawn at an arbitrarily given angle to the given line. For, the original $2n$-line problem literally involves not distances to the given lines, but lengths of straight lines drawn at given angles to the given lines. For the methods of Descartes, the distinction is trivial. For Apollonius, the distinction would seem not to be trivial.

The question remains: If Descartes can express the solution of a locus problem in terms that would make sense to Apollonius or Pappus, would the ancient mathematician accept Descartes's *proof,* a proof that involves algebraic manipulations of symbols?

8 A book from the 1990s

In 2006 in Ankara, with two colleagues, I taught a first-year, first-semester undergraduate analytic geometry course from a locally published text that was undated, but had apparently been produced in 1994 [24]. The preface of that text begins:

> This book is meant as a basic text book for a course in Analytic Geometry.
>
> Throughout the book, the connections and interrelations between algebra and geometry are emphasized. the notions of Linear Algebra are introduced and applied simultaneously with more traditional topics of Analytic Geometry. Some of the notions of Linear Algebra are used without mentioning them explicitly.

The preface continues with brief descriptions of the eight chapters and two appendices, and it concludes with acknowledgements. The text's Chapter 1, "Fundamental Principle of Analytic Geometry", has five sections:

1. Set Theory
2. Relations
3. Functions
4. Families of Sets
5. Fundamental Principle of Analytic Geometry

Thus the book appears more sophisticated than the 1949 book discussed in Chapter 5. Possibly this shows the influence of the intervening New Math in the US, if the text draws on American sources; but here I am only speculating. The author's acknowledgements include no written sources, and the book has no bibliography. The introduction to Chapter 1 reads:

> Analytic Geometry is a branch of mathematics which studies geometry through the use of algebra. It was *Rene Descartes* (1596–1650) who introduced the subject for the first time. Analytic geometry is based on the observation that there is a one-to-one

> correspondence between the points of a straight line and the real numbers (see §5). This fact is used to introduce coordinate systems in the plane or in three space, so that a geometric object can be viewed as a set of pairs of real numbers or as a set of triples of real numbers.
>
> In this chapter, we list notations, review set theoretic notions and give the fundamental principle of analytic geometry.

The reference to Descartes is too vague to be meaningful. Descartes does not *observe,* but he tacitly *assumes,* that there is a one-to-one correspondence between lengths and *positive* numbers. He assumes too that numbers can be multiplied by one another; but in case there is any question about this assumption, he *proves* that this multiplication is induced by a geometrically meaningful notion. His proof is discussed further below in Chapter 9.

As spelled out on pages 15 and 16 in the book under review, the **Fundamental Principle of Analytic Geometry** is that for every straight line ℓ there is a function P from $\mathbb{R}$ to ℓ such that:

a) $P(\mathsf{0}) \neq P(\mathsf{1})$;
b) for every positive integer n, the points $P(\pm n)$ are n times as far away from $P(\mathsf{0})$ as $P(\mathsf{1})$ is, and are on the same and opposite sides of $P(\mathsf{0})$ respectively;
c) similarly for the points $P(\pm k)$ and $P(k/n)$, when k is also a positive integer;
d) if $x < y$, then the direction from $P(x)$ to $P(y)$ is the same as from $P(\mathsf{0})$ to $P(\mathsf{1})$.

It follows that any choice of distinct points $P(\mathsf{0})$ and $P(\mathsf{1})$ uniquely determines such a function P.

In a more rudimentary form, this Fundamental Principle is called the **Cantor–Dedekind Axiom** on Wikipedia.[1] I would analyze this Principle or Axiom into two parts:

1. For a point O on a straight line, for one of the two sides of O, the line has the structure of an ordered (abelian) group in

[1] Article of that name accessed October 17, 2013. It says "the phrase *Cantor–Dedekind axiom* has been used to describe the thesis that the real numbers are order-isomorphic to the linear continuum of geometry." No preservation of *algebraic* structure is discussed.

which the chosen point is the neutral element and the positive elements are on the chosen side of O. If A and B are arbitrary points on the line, then $A+B$ is that point C such that the segments OB and AC are congruent and C is on the same side of A that B is of O.
2. If a particular point U of the line is chosen on the chosen side of O, then there is an isomorphism P from the ordered group of real numbers to the ordered group of the line in which $P(1)=U$.

The first part here defines addition of points compatibly with the addition of segments defined in the text of Nelson *& al.* discussed in Chapter 5 above. The second part establishes *continuity* of straight lines in the sense of Dedekind discussed in Chapter 3 above.

In fact this two-part formulation of the Fundamental Principle is strictly stronger than what the text gives. By the text version, the map P is not a group homomorphism; only its restriction to $\mathbb{Q}$ is a group homomorphism. As noted in Chapter 2 above, by Dedekind's construction, $\mathbb{R}$ is initially obtained from $\mathbb{Q}$ as a linear order alone; it must be *proved* to have field-operations extending those of $\mathbb{Q}$. By the second part of the two-part formulation of the Fundamental Principle, *addition* on $\mathbb{R}$ is geometrically meaningful. This is left out of the Principle as formulated in the text under review.

As Dedekind observes, and as was repeated in Chapter 3 above, continuity is not necessary for doing geometry. Thus the so-called Fundamental Principle is not necessary. It is not even *sufficient* for doing geometry; for it provides no clue about what happens away from a given straight line. The Principle holds for every Riemannian manifold with no closed geodesics. In such a manifold, a chosen point on a geodesic and a chosen direction along the geodesic determine the structure of an ordered group on the geodesic, and this ordered group is isomorphic with $\mathbb{R}$ as an ordered group. The bijection from $\mathbb{R}$ to the geodesic induces a multiplication on the geodesic; but this multiplication is not generally of significance within the manifold. It is of significance in a *Euclidean* manifold, where we have an equation between the product of two lengths and the area of a rectangle whose dimensions are those lengths. This equation is fundamental to analytic geometry; but the so-called Fundamental Principle of Analytic Geometry does not give it to us.

The first theorem in the book under review is that the usual formula for the distance between two points is correct. The proof appeals to the Pythagorean Theorem without further explanation. This is a failure of rigor. Proposition I.47 of Euclid's *Elements* gives us the Pythagorean Theorem as an equation of certain linear combinations of *areas.* To do analytic geometry, we need to be able to understand this as an equation of certain polynomial functions of *lengths.* If all lengths are commensurable, this is easy. Since not all lengths are commensurable, more work is needed, which will be discussed in the next chapter.

The second theorem in the book under review is that every straight line is defined by a linear equation as follows.

1. A vertical straight line is defined by an equation $x = a$.
2. A horizontal straight line is defined by an equation $y = b$.
3. If the line is inclined, then any two points of the line give the same slope for the line, and so the line is defined by an equation $y = m(x - x_1) + y_1$.

This last conclusion is justified by similarity of triangles. The possibility of distinguishing straight lines as vertical, horizontal, or oblique is asserted without explanation. The meaning of straightness is not discussed.

Each of the equations that have been found for a line can be put into the form $Ax + By + C = 0$.[2] The third theorem of the text is the converse. If one of A and B is not 0,[3] then the equation $Ax + By + C = 0$ defines

(1) the vertical line defined by $x = -C/A$, if $B = 0$;
(2) the straight line through $(0, -C/B)$ having slope $-A/B$, if $B \neq 0$.

[2]According to the text, the equations $x = a$, $y = b$, and $y = m(x - x_1) + y_1$ are called *defining equations* of the corresponding lines, and an equation of the form $Ax + By + C = 0$ is called a linear equation. The second theorem is given as, "The defining equation of any straight line is a linear equation." Taken literally, this is trivial.

[3]This condition is omitted from the text.

This assumes that a point and a slope determine a line.

It seems to me that these first three theorems are founded on notions from high-school mathematics, but add no rigor to these notions. It would be more honest to say something like,

> As we know from high school, straight lines are just graphs of equations of the form $Ax + By + C = 0$, where at least one of A and B is not 0...

To write this out as formal numbered theorems, with proofs labelled as "Proof" and ended with boxes □—this is a failure of rigor, unless the axioms that the proofs rely on are made explicit.[4] We have already seen that the Fundamental Principle of Analytic Geometry is an insufficient axiomatic foundation.

I argued in Chapter 3 above that Euclid's Proposition I.4, "Side-Angle-Side," is reasonably treated as a theorem, rather than a postulate, even though it relies on no postulates. But Euclid is not working within a formal system. He has no such notion. Today we have the notion, and a textbook of mathematics ought to give at least a nod to the reader who is familiar with the notion.

Descartes is more rigorous than the book considered here, even though he does not write out any theorems as such. It is clear that his logical basis is Euclidean geometry as used by the ancient Greek mathematicians.

[4]In the book under review, the first three theorems get the numbers 2.1.1, 2.2.1, and 2.2.2. The first and third have proofs ended with boxes. The second is preceded by two pages of discussion and diagrams, followed by "We have thus proved".

9 Geometry to algebra

To consider the matter of rigor in more detail, I propose to compare the so-called Fundamental Principle of Analytic Geometry in the previous chapter with the axioms of David Hilbert [22]. The latter are given in five groups. For plane geometry, the axioms can be summarized as follows.

I. Connection.

1–2. Two distinct points lie on a unique straight line.[1]

7. A line contains at least two points.

II. Order.

1–4. The points of a straight line are densely linearly ordered without extrema.

5 (Pasch's Axiom). A straight line intersecting one side of a triangle intersects one of the other two sides or meets their common vertex.

III. Parallels (Euclid's Axiom).

Through a given point, exactly one parallel to a given straight line can be drawn.

IV. Congruence.

1. Every segment can be uniquely *laid off* upon a given side of a given point of a given straight line.

2–3. Congruence of segments is transitive and additive.

[1] Hilbert writes this axiom as two axioms, although the distinction between the two is obscure. If the two axioms are to be logically independent from one another, then I think the first axiom should be understood as being that for any two distinct points A and B, there is at least one straight line, called AB or BA, that contains them. Then the second axiom is that, if A, B, and C are distinct points, and $AB = AC$, then $AB = BC$. It would then be a *theorem* that two distinct points lie on a *unique* straight line. However, Hilbert does not mention such a theorem. Perhaps he considers it to be an immediate consequence of his first axiom. But in that case, his second axiom would appear to be redundant.

4. Every angle can be uniquely *laid off* upon a given side of a given ray.[2]
5. Congruence of angles is transitive.
6. "Side-Angle-Side" (Euclid's Proposition I.4).

V. Continuity (Archimedean Axiom).

Some multiple of one segment exceeds another.

Hilbert gives an additional **Axiom of Completeness,** that no larger system satisfies the axioms.

In the two-part formulation of the Fundamental Principle of Analytic Geometry given in Chapter 8 previous, the first part is equivalent to Hilbert's Order and Congruence Axioms, as restricted to a single straight line.

Granted that Hilbert's axioms allow the construction of an ordered field K as discussed below, Pasch's axiom ensures that "space" has at most two dimensions. The Completeness Axiom then ensures that space has exactly two dimensions. Then the Completeness and Continuity Axioms together ensure that the ordered field K is $\mathbb{R}$. Indeed, these two axioms, in the presence of the others, are equivalent to the second part of the Fundamental Principle of Analytic Geometry.

Hilbert shows that the Axiom of Parallels and the "Side-Angle-Side" axiom respectively are independent from all of the other axioms.[3] In particular then, the Fundamental Principle of Algebraic Geometry is not sufficient for doing geometry. We have already cited Dedekind to the effect that continuity is not necessary for doing geometry.

What *is* needed for doing *analytic* geometry is what Descartes observes: Given an ordered-group isomorphism P from a field K to a straight line with a distinguished point and direction, we must be able to obtain $P(x \cdot y)$ from $P(x)$ and $P(y)$ in a geometrically meaningful way. For Descartes, this grounds algebra in geometry and so makes *algebra* rigorous.

---

[2] Hilbert (or his translator) says "half-ray".

[3] Strictly, Hilbert leaves the Completeness Axiom out of his arguments; but leaving it in would not affect his independence proofs.

Having picked a unit length, Descartes defines a multiplication of lengths by means of the theory of proportion—presumably the theory of Book V of the *Elements.* In particular, Descartes implicitly uses Euclid's Proposition VI.2, that a straight line parallel to the base of a triangle divides the sides proportionally. If one side is divided into parts of lengths 1 and a, then the other side is divided into parts of lengths b and ab for some b, as in Figure 9.1; and a and

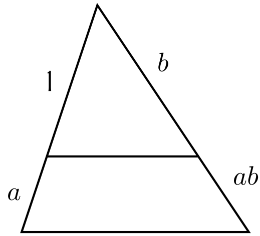


Figure 9.1: Multiplication of lengths

b can be chosen in advance.

As developed in Euclid, the theory of proportion uses the Archimedean Axiom. Hilbert shows how to avoid using this Axiom, but the arguments are somewhat complicated. Hartshorne [16] has a more streamlined approach, using properties of circles in Book III of the *Elements.*

In fact, Euclid's Book I alone provides a sufficient basis for defining multiplication. It was suggested in Chapter 3 above that Proposition 45 is the climax of that book. It was observed in Chapter 6 that Proposition 44 is a lemma for this proposition. This lemma in turn relies heavily on Proposition 43:

> In any parallelogram the complements of parallelograms about the diameter are equal to one another.

In particular, in rectangle ΑΒΓΔ in Figure 9.2, the diagonal ΑΓ is taken, and rectangles ΕΘ and ΗΖ are drawn sharing this diagonal. Then the complementary rectangles ΒΚ and ΚΔ are equal to one another, because they are the same linear combinations of triangles that are respectively equal to one another.

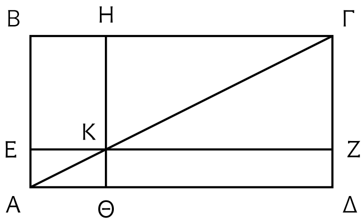


Figure 9.2: Euclid's Proposition I.43

Now we can define the product of two lengths a and b as in Figure 9.3, so that ab is the width of a rectangle of unit height that is

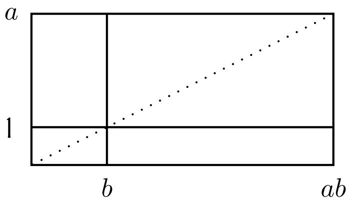


Figure 9.3: Multiplication of lengths

equal to a rectangle of dimensions a and b. Given the theorem (easily proved) that all rectangles of the same dimensions are equal, multiplication is automatically commutative. Easily too, it distributes over addition, and there are multiplicative inverses. Associativity takes a little more work. In Figure 9.4, by definition of ab, cb, and $a(cb)$, we have

$$(*) \qquad \begin{cases} A + B = E + F + H + K, \\ C = G, \\ A = D + E + G + H. \end{cases}$$

Also $a(cb) = c(ab)$ if and only if

$$C + D + E = K.$$

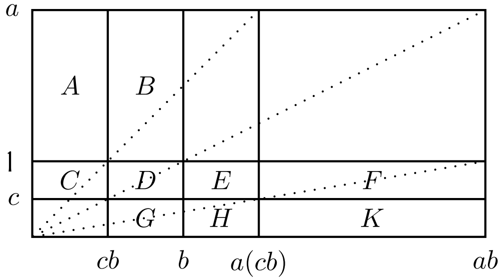


Figure 9.4: Associativity of multiplication of lengths

From $(*)$ we compute

$$D + C + B = F + K.$$

We finish by noting that, by Euclid's Proposition I.43,

$$B = E + F.$$

Thus we can establish, by geometric means alone, that lengths are the positive elements of an *ordered field. This* is what makes analytic geometry possible.

Descartes apparently does not recognize any need to establish commutativity, distributivity, and associativity of multiplication. Still he should be credited with the observation that a geometric definition of multiplication of lengths is what is needed for the application of algebra to geometry. It is a shame that textbooks should cite Descartes as the creator of analytic geometry when his fundamental insight is forgotten.

10 Algebra to geometry

One can work the other way. One can start with an ordered field K, and one can interpret the product $K \times K$ as a Euclidean plane. Hilbert sketches the argument in order to prove that his axioms for such a plane are consistent [22, §9, pp. 17-8].

In fact an arbitrary ordered field is not sufficient. To provide a model for Hilbert's axioms, K must also be **Pythagorean,** that is, closed under the operation $x \mapsto \sqrt{1 + x^2}$. This allows hypotenuses of right triangles to have lengths in the field.

For Euclid's postulates, K must be **Euclidean,** that is, closed under taking square roots of all positive elements. Indeed, Descartes defines square roots geometrically so that $\sqrt{a}$ is, in effect, the ordinate corresponding to the abscissa 1 in a circle of diameter $1 + a$. Hilbert shows [22, §37, pp. 73-4] that the Euclidean condition is strictly stronger than the Pythagorean condition.[1] The Pythagorean closure of $\mathbb{Q}$ contains the conjugates of all of its elements. But the conjugate of a positive element need not be positive. For example, $\sqrt{2} - 1$ is positive, but its conjugate $-\sqrt{2} - 1$ is not. Thus $\sqrt{2} - 1$ has no square root in the Pythagorean closure of $\mathbb{Q}$.

It is perhaps odd that Hilbert should feel the need to prove his axioms consistent. One might consider them as self-evidently consistent. Hilbert bases his consistency argument on the existence of a Pythagorean ordered field. One might argue that we believe such fields exist only because of *geometric* demonstrations like Descartes's as discussed in Chapter 9 previous.

On the other hand, we can obtain the ordered field $\mathbb{Q}$ without geometry, and Dedekind shows how to do the same for $\mathbb{R}$. The student may know something about $\mathbb{R}$ from calculus class. Then the student also knows about the product $\mathbb{R} \times \mathbb{R}$ from calculus, but

[1] Hilbert does not use this terminology.

mainly as the setting for graphs of functions. With this background, how can the student recover Euclidean geometry?

By means of the structure of $\mathbb{R}$ as an ordered field, we can give $\mathbb{R} \times \mathbb{R}$ the structure of an inner product space:

1. $\mathbb{R} \times \mathbb{R}$ has the abelian group structure induced from $\mathbb{R}$ itself.
2. By the standard multiplication, $\mathbb{R}$ acts on $\mathbb{R}\times\mathbb{R}$ as a field, that is, $\mathbb{R}$ embeds in the (unital associative) ring of endomorphisms of $\mathbb{R} \times \mathbb{R}$ as an abelian group. Thus $\mathbb{R} \times \mathbb{R}$ is a vector space over $\mathbb{R}$.
3. The function $(\boldsymbol{x}, \boldsymbol{y}) \mapsto \boldsymbol{x} \cdot \boldsymbol{y}$ from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$ given by
$$\boldsymbol{x} \cdot \boldsymbol{y} = x_0 y_0 + x_1 y_1$$
is a real inner product function, that is, a positive-definite, symmetric, bilinear function.
4. Hence there is a norm function $\boldsymbol{x} \mapsto |\boldsymbol{x}|$ given by
$$|\boldsymbol{x}| = \sqrt{\boldsymbol{x} \cdot \boldsymbol{x}}.$$

We declare that the **distance** between two points $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathbb{R} \times \mathbb{R}$ is $|\boldsymbol{b} - \boldsymbol{a}|$. We must check that distance has the geometrical properties that we expect. The most basic of these properties are found in the definition of a metric. Easily, the distance function is symmetric. Also, the distance between distinct points is positive, while the distance between identical points is zero. We can establish the *Triangle Inequality* by means of the *Cauchy–Schwartz Inequality.*

To this end, we first define two elements $\boldsymbol{a}$ and $\boldsymbol{b}$ of $\mathbb{R} \times \mathbb{R}$ to be **parallel,** writing
$$\boldsymbol{a} \parallel \boldsymbol{b},$$
if the equation
$$x\boldsymbol{a} + y\boldsymbol{b} = \mathbf{0}$$
has a nonzero solution. Given arbitrary $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathbb{R} \times \mathbb{R}$, where $\boldsymbol{a} \neq \mathbf{0}$, we have the identity
$$\sum_{i<2}(a_i x + b_i)^2 = |\boldsymbol{a}|^2 x^2 + 2(\boldsymbol{a} \cdot \boldsymbol{b})x + |\boldsymbol{b}|^2.$$

If $\boldsymbol{a} \parallel \boldsymbol{b}$, there is exactly one zero to this quadratic polynomial, and so the discriminant is zero. Otherwise there are no zeros, so the discriminant is negative. Therefore we have the **Cauchy–Schwartz Inequality**

$$|\boldsymbol{a} \cdot \boldsymbol{b}| \leqslant |\boldsymbol{a}| \cdot |\boldsymbol{b}|,$$

with equality if and only if $\boldsymbol{a} \parallel \boldsymbol{b}$. Hence

$$(*) \qquad \begin{cases} |\boldsymbol{a}+\boldsymbol{b}|^2 = (\boldsymbol{a}+\boldsymbol{b}) \cdot (\boldsymbol{a}+\boldsymbol{b}) \\ \qquad\qquad = |\boldsymbol{a}|^2 + 2\boldsymbol{a} \cdot \boldsymbol{b} + |\boldsymbol{b}|^2 \\ \qquad\qquad \leqslant |\boldsymbol{a}|^2 + 2|\boldsymbol{a}| \cdot |\boldsymbol{b}| + |\boldsymbol{b}|^2 \\ \qquad\qquad = (|\boldsymbol{a}| + |\boldsymbol{b}|)^2, \end{cases}$$

with equality if and only if $\boldsymbol{a} \parallel \boldsymbol{b}$ and $\boldsymbol{a} \cdot \boldsymbol{b} \geqslant 0$. This condition is that $\boldsymbol{a}$ and $\boldsymbol{b}$ are in the **same direction.** We obtain the **Triangle Inequality**

$$|\boldsymbol{a}+\boldsymbol{b}| \leqslant |\boldsymbol{a}| + |\boldsymbol{b}|,$$

with equality if and only if $\boldsymbol{a}$ and $\boldsymbol{b}$ are in the same direction.

We can now define the **line segment** with distinct endpoints $\boldsymbol{a}$ and $\boldsymbol{b}$ to be the set of points $\boldsymbol{x}$ such that

$$|\boldsymbol{b}-\boldsymbol{a}| = |\boldsymbol{b}-\boldsymbol{x}| + |\boldsymbol{x}-\boldsymbol{a}|.$$

This is just the set of $\boldsymbol{x}$ such that $\boldsymbol{b}-\boldsymbol{x}$ and $\boldsymbol{x}-\boldsymbol{a}$ are in the same direction. We may say in this case that $\boldsymbol{x}$ is **between** $\boldsymbol{a}$ **and** $\boldsymbol{b}$. We should note that this definition is symmetric in $\boldsymbol{a}$ and $\boldsymbol{b}$. The segment can be denoted indifferently by $\boldsymbol{ab}$ or $\boldsymbol{ba}$. The distance between $\boldsymbol{a}$ and $\boldsymbol{b}$ is the **length** of this segment. Two segments are **equal** if they have the same length.

The **(straight) line** containing distinct points $\boldsymbol{a}$ and $\boldsymbol{b}$ consists of those $\boldsymbol{x}$ such that

$$\boldsymbol{x}-\boldsymbol{a} \parallel \boldsymbol{b}-\boldsymbol{a}.$$

The points $\boldsymbol{a}$ and $\boldsymbol{b}$ **determine** this line. Any two distinct points of the line determine it.

The **circle** with **center** $\boldsymbol{a}$ passing through $\boldsymbol{b}$ consists of all $\boldsymbol{x}$ such that
$$|\boldsymbol{x}-\boldsymbol{a}| = |\boldsymbol{b}-\boldsymbol{a}|.$$
The **radius** of this circle is $|\boldsymbol{b}-\boldsymbol{a}|$.

It appears we now have Euclid's first three postulates. But now is when the accusation that Euclid uses hidden assumptions becomes meaningful. In the present context, we do not automatically have equilateral triangles by Euclid's Proposition I.1. In the proof as described through Figure 3.1 (in Chapter 3) above, one must check that the two circles do indeed intersect. Here one would use that $\mathbb{R}$ contains $\sqrt{3}$.

For Proposition I.4, we need the notion of **angle.** We can define it as the union of two line segments that share a common endpoint, provided that the the other endpoints are not collinear with the common endpoint. The union of the segments $\boldsymbol{ab}$ and $\boldsymbol{ac}$ can be denoted indifferently by $\boldsymbol{bac}$ or $\boldsymbol{cab}$. If we have the cosine function cos from the interval $(0, \pi)$ to $(-1, 1)$, along with its inverse arccos, defined analytically by power series, we can define the **measure** of the angle $\boldsymbol{bac}$ as
$$\arccos \frac{(\boldsymbol{c}-\boldsymbol{a}) \cdot (\boldsymbol{b}-\boldsymbol{a})}{|\boldsymbol{c}-\boldsymbol{a}| \cdot |\boldsymbol{b}-\boldsymbol{a}|}.$$
Then two angles are **equal** if they have the same measure. If $\boldsymbol{d}$ and $\boldsymbol{e}$ are distinct from $\boldsymbol{a}$, and $\boldsymbol{b}-\boldsymbol{a}$ and $\boldsymbol{d}-\boldsymbol{a}$ are in the same direction, and likewise for $\boldsymbol{c}-\boldsymbol{a}$ and $\boldsymbol{e}-\boldsymbol{a}$, then angles $\boldsymbol{bac}$ and $\boldsymbol{dae}$ are equal.

Our earlier computations now give us the **Law of Cosines:** If $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ are three noncollinear points, so that they are vertices of a triangle, and if the measure of angle $\boldsymbol{bac}$ is θ, then by $(*)$ we have
$$|\boldsymbol{b}-\boldsymbol{c}|^2 = |\boldsymbol{b}-\boldsymbol{a}|^2 + |\boldsymbol{a}-\boldsymbol{c}|^2 - 2|\boldsymbol{b}-\boldsymbol{a}| \cdot |\boldsymbol{a}-\boldsymbol{c}| \cdot \cos\theta.$$
Hence the lengths of the sides $\boldsymbol{ab}$ and $\boldsymbol{ac}$ and the measure of the angle they include determine the length of the opposite side $\boldsymbol{bc}$. Conversely, the lengths of all three sides determine the measure of each angle. So we have Euclid's I.4 and also I.8 ("Side-Side-Side").

We have relied on the cosine function and its inverse, although, being defined by power series, they are not algebraic: their use here assumes that our ordered field is not arbitrary, but is indeed $\mathbb{R}$. However, we need not actually find measures of angles. Two angles are equal if and only if their cosines are equal; and these cosines are defined algebraically from the sides of the angles.

Hilbert seems to allude to a different approach, apparently the approach pioneered by Felix Klein [38, p. 138]. We can just declare any two subsets of $\mathbb{R} \times \mathbb{R}$ to be **congruent** if one can be carried to the other by a **translation**

$$\boldsymbol{x} \mapsto \boldsymbol{x} + \boldsymbol{a}$$

followed by a **rotation**

$$(x, y) \mapsto (x \cos\theta - y \sin\theta, x \sin\theta + y \cos\theta)$$

or a **reflection**

$$(x, y) \mapsto (x \cos\theta + y \sin\theta, x \sin\theta - y \cos\theta).$$

Then two line segments will be equal if and only if they are congruent. The same will go for angles, provided we define these as unions of *rays* rather than segments.

Bibliography

[1] Apollonius of Perga. *Apollonius of Perga: Treatise on Conic Sections.* University Press, Cambridge, UK, 1896. Edited by T. L. Heath in modern notation, with introductions including an essay on the earlier history of the subject.

[2] Apollonius of Perga. *Apollonii Pergaei qvae Graece exstant cvm commentariis antiqvis*, volume I. Teubner, 1974. Edidit et Latine interpretatvs est I. L. Heiberg.

[3] Apollonius of Perga. *Conics. Books I–III.* Green Lion Press, Santa Fe, NM, revised edition, 1998. Translated and with a note and an appendix by R. Catesby Taliaferro, With a preface by Dana Densmore and William H. Donahue, an introduction by Harvey Flaumenhaft, and diagrams by Donahue, Edited by Densmore.

[4] Tom M. Apostol. *Mathematical analysis.* Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., second edition, 1974.

[5] Archimedes. *The two books* On the sphere and the cylinder, volume I of *The works of Archimedes.* Cambridge University Press, Cambridge, 2004. Translated into English, together with Eutocius' commentaries, with commentary, and critical edition of the diagrams, by Reviel Netz.

[6] Carl B. Boyer. *A History of Mathematics.* John Wiley & Sons, New York, 1968.

[7] Richard Dedekind. *Essays on the theory of numbers. I: Continuity and irrational numbers. II: The nature and meaning of numbers.* authorized translation by Wooster Woodruff Beman. Dover Publications Inc., New York, 1963.

[8] René Descartes. *The Geometry of René Descartes.* Dover Publications, Inc., New York, 1954. Translated from the French and Latin by David Eugene Smith and Marcia L. Latham, with a facsimile of the first edition of 1637.

[9] René Descartes. *The Philosophical Writings of Descartes*, volume I. Cambridge University Press, 1985. translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch.

[10] Jean Dubois, Henri Mitterand, and Albert Dauzat. *Dictionnaire d'Étymologie.* Larousse, Paris, 2001. First edition 1964.

[11] Euclid. *The thirteen books of Euclid's Elements translated from the text of Heiberg. Vol. I: Introduction and Books I, II. Vol. II: Books III–IX. Vol. III: Books X–XIII and Appendix.* Dover Publications Inc., New York, 1956. Translated with introduction and commentary by Thomas L. Heath, 2nd ed.

[12] Euclid. *The Bones.* Green Lion Press, Santa Fe, NM, 2002. A handy where-to-find-it pocket reference companion to Euclid's *Elements*.

[13] Euclid. *Euclid's Elements.* Green Lion Press, Santa Fe, NM, 2002. All thirteen books complete in one volume. The Thomas L. Heath translation, edited by Dana Densmore.

[14] David Fowler. *The mathematics of Plato's academy.* Clarendon Press, Oxford, second edition, 1999. A new reconstruction.

[15] Carl Friedrich Gauss. *Disquisitiones Arithmeticae.* Springer-Verlag, New York, 1986. Translated into English by Arthur A. Clarke, revised by William C. Waterhouse.

[16] Robin Hartshorne. *Geometry: Euclid and beyond.* Undergraduate Texts in Mathematics. Springer-Verlag, New York, 2000.

[17] T. L. Heath. *The Works of Archimedes: Edited in Modern Notation With Introductory Chapters.* University Press, Cambridge, UK, 1897.

[18] Thomas Heath. *A history of Greek mathematics. Vol. II.* Dover Publications Inc., New York, 1981. From Aristarchus to Diophantus, Corrected reprint of the 1921 original.

[19] Thomas L. Heath, editor. *The* Method *of Archimedes Recently Discovered by Heiberg: A supplement to* The Works of Archimedes *1897.* University Press, Cambridge, UK, 1912.

[20] Leon Henkin. On mathematical induction. *Amer. Math. Monthly*, 67:323–338, 1960.

[21] Herodotus. *The Persian Wars, Books I–II*, volume 117 of *Loeb Classical Library.* Harvard University Press, Cambridge, Massachusetts and London, England, 2004. Translation by A. D. Godley; first published 1920; revised, 1926.

[22] David Hilbert. *The foundations of geometry.* Authorized translation by E. J. Townsend. Reprint edition. The Open Court Publishing Co., La Salle, Ill., 1959. Project Gutenberg edition released December 23, 2005 (`www.gutenberg.net`).

[23] T. F. Hoad, editor. *The Concise Oxford Dictionary of English Etymology.* Oxford University Press, Oxford and New York, 1986. Reissued in new covers, 1996.

[24] H. I. Karakaş. *Analytic Geometry.* M $\oplus$ V [Matematik Vakfı], [Ankara], n.d. [1994].

[25] Morris Kline. *Mathematical thought from ancient to modern times.* Oxford University Press, New York, 1972.

[26] Edmund Landau. *Foundations of Analysis. The Arithmetic of Whole, Rational, Irrational and Complex Numbers.* Chelsea Publishing Company, New York, N.Y., third edition, 1966. translated by F. Steinhardt; first edition 1951; first German publication, 1929.

[27] Henry George Liddell and Robert Scott. *A Greek-English Lexicon.* Clarendon Press, Oxford, 1996. revised and augmented

throughout by Sir Henry Stuart Jones, with the assistance of Roderick McKenzie and with the cooperation of many scholars. With a revised supplement.

[28] William Morris, editor. *The Grolier International Dictionary.* Grolier Inc., Danbury, Connecticut, 1981. two volumes; appears to be the *American Heritage Dictionary* in a different cover.

[29] James Morwood, editor. *The Pocket Oxford Latin Dictionary.* Oxford University Press, 1995. First edition published 1913 by Routledge & Kegan Paul.

[30] Murray et al., editors. *The Compact Edition of the Oxford English Dictionary.* Oxford University Press, 1973.

[31] Alfred L. Nelson, Karl W. Folley, and William M. Borgman. *Analytic Geometry.* The Ronald Press Company, New York, 1949.

[32] Pappus. *Pappus Alexandrini Collectionis Quae Supersunt*, volume II. Weidmann, Berlin, 1877. E libris manu scriptis edidit, Latina interpretatione et commentariis instruxit Fridericus Hultsch.

[33] David Pierce. Induction and recursion. *The De Morgan Journal*, 2(1):99–125, 2012. `http://education.lms.ac.uk/2012/04/david-pierce-induction-and-recursion/`.

[34] Proclus. *Procli Diadochi in primum Euclidis Elementorum librum commentarii.* Bibliotheca scriptorum Graecorum et Romanorum Teubneriana. In aedibus B. G. Teubneri, 1873. Ex recognitione Godofredi Friedlein.

[35] Procopius. *On Buildings*, volume 343 of *Loeb Classical Library.* Harvard University Press, Cambridge, Massachusetts, and London, England, revised edition, 1954. with an English translation by H. B. Dewing.

[36] Abraham Robinson. *Non-standard analysis.* Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1996. Reprint of the second (1974) edition, With a foreword by Wilhelmus A. J. Luxemburg.

[37] Lucio Russo. *The forgotten revolution.* Springer-Verlag, Berlin, 2004. How science was born in 300 BC and why it had to be reborn, translated from the 1996 Italian original by Silvio Levy.

[38] R. W. Sharpe. *Differential geometry*, volume 166 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1997. Cartan's generalization of Klein's Erlangen program, With a foreword by S. S. Chern.

[39] Walter W. Skeat. *A Concise Etymological Dictionary of the English Language.* Perigee Books, New York, 1980. First edition 1882; original date of this edition not given.

[40] Michael Spivak. *Calculus. 2nd ed.* Berkeley, California: Publish or Perish, Inc. XIII, 647 pp., 1980.

[41] Ivor Thomas, editor. *Selections illustrating the history of Greek mathematics. Vol. I. From Thales to Euclid*, volume 335 of *Loeb Classical Library.* Harvard University Press, Cambridge, Mass., 1951. With an English translation by the editor.

[42] Ivor Thomas, editor. *Selections illustrating the history of Greek mathematics. Vol. II. From Aristarchus to Pappus*, volume 362 of *Loeb Classical Library.* Harvard University Press, Cambridge, Mass, 1951. With an English translation by the editor.

[43] John von Neumann. On the introduction of transfinite numbers (1923). In Jean van Heijenoort, editor, *From Frege to Gödel: A source book in mathematical logic, 1879–1931*, pages 346–354. Harvard University Press, Cambridge, MA, 2002.