# 1

# Introduction

A famous anecdote about Gauss would have him measuring the angles of the great triangle formed by the mountain peaks of Inselberg, Brocken, and Hohenhagen, while looking for evidence that the geometry of space is non-Euclidean. Whether true or simply a myth, this oft told story emphasizes the tug of war that has existed between Plato, Kant, Poincaré and other idealists, *a priorists* and conventionalists, and those who have insisted that the geometrical structure of our world is an empirical matter of fact, to be decided by experience.

For many years the debate was confined to the question of curvature. Non-Euclidean geometries were initially believed to be impossible, and even after their discovery, they were still considered inappropriate for describing the geometry of three-dimensional space. Flat (Euclidean) geometry, on the other hand, was always the common view, chosen by default as ideal, or *a priori*, or simply more convenient.

The history of science has taught us otherwise. As Einstein once said [Ein29], "People slowly accustomed themselves to the idea that the physical states of space itself were the final physical reality." The geometry of this reality turned out to be constrained by the distribution of matter, hence by empirical, contingent, matters of fact. Moreover, the general theory of relativity that describes it has turned physical objects (matter fields) into geometrical objects, and geometrical objects (the metric and the affine connection) into physical objects.

In this book we shall focus on another feature of geometry that exemplifies this tug of war between the *a priorists* and the empiricists. Rather than curvature, we shall be dealing with the structure of the line segment to which one-dimensional space is believed to be isomorphic. To the naked eye this one-dimensional space looks continuous, much as a film made of 24 frames per second does, but is it truly so at its most fundamental level?

Of course, there is no way to answer this question by direct observation. As humans, we are bounded by finite resolution capabilities; no matter how fine grained our “rulers” may be, their resolution can never directly reveal the putative fundamental continuum of the physical line segment. This line segment may be fundamentally discrete (and may appear discrete to our most refined “rulers”), or it may be fundamentally continuous (but still appear discrete to our most refined “rulers”). And yet, without evidence to the contrary, and consistent with our coarse grained (naked eye) observations, we indirectly *infer* a putative, continuous structure from the successful applicability of continuum geometry, the calculus, and the field concept with its continuous symmetries, to physical phenomena *in* space. This extrapolation is based on three premises: the mathematical structure with which we describe phenomena in physical space is continuous; the theories relying on this structure have been verified down to a scale of $10^{-16}$ cm; and we strongly believe that physical space is isomorphic to this structure, regardless of our actual finite resolution capabilities. Such an inference leaves plenty of room for debate on the possibility of alternative structures.

Our journey through the landscape of these alternative structures begins in Chapter 2, with the history and the philosophy of mathematics, where ingenious arguments have been constructed for (and against) spatial discreteness, as well as for (and against) its rival, the continuum. These arguments aim to establish that only one of these notions is logically consistent as a basis for geometry, and moreover, that only one of them could represent faithfully the structure of physical space.

The scope of these arguments is overwhelmingly vast. They range from topology to geometry, and from number theory to the theory of computation. But they all fail to establish their conclusion; there is nothing inconsistent in a geometry based on a discrete line segment (or on a continuous line segment, for that matter), and there is more than one way to describe the physical world. Notwithstanding the ubiquity of the calculus, spacetime may be discrete at the most fundamental level, and only appear continuous (that is, be faithfully represented by continuous geometry) at lower resolutions. Even more important, these arguments demonstrate the disagreement that prevails already with regard to the question of whether the character of the line segment that faithfully represents the physical world can be decided empirically. Throughout this book we shall clarify what it takes to answer this question.

More *a priori* arguments against finitism – treated here as a general philosophical thesis – are presented (and dismantled) in Chapter 3. These arguments target the putative philosophical price that finitism may carry, saddling it with metaphysical and epistemological commitments that appear to have dire consequences. For example, finitism may blur the distinction between metaphysics and epistemology by subordinating what *is* to what we *know*. All by itself this may not seem a serious allegation, but in the context of interpreting, say, the notion of probability in quantum mechanics, such blurring makes it difficult to distinguish quantum from classical probabilities.

Difficult, but not impossible: one can still identify the difference between classical and quantum probabilities in the structure of the respective probability spaces, i.e., whether this structure is Boolean or non-Boolean. This distinction sees quantum interference not as evidence for some metaphysical difference between classical and quantum "realities," but rather as representing, under several plausible physical assumptions, a quantitative difference in measure between the two. Remarkably, such a finitist and objective interpretation of quantum probability as a limit on measurement resolution is also consistent with the way the notion of fundamental length was introduced into modern physics, and it can be used to restore credibility to the thought experiments that motivated it, and to their putative implications.

Another angle on this tension between ontology and epistemology is that finitism may spill the baby with the tub water, as it threatens to turn the question of spatial discreteness into a metaphysical, undecidable question. If what *is* is dictated by what we *know*, then, or so the story goes, given our finite resolution capabilities, we could never decide whether the world is truly discrete or continuous. But here also, that the question of spatial discreteness is undecidable does not follow from the operational methodology that regards measurement results (which are always finite) as primitive. To see this, an operational argument that goes back to Einstein about the primacy of geometrical notions is shown to be distinct from the views of his contemporaries on the conventionality or the *a priori* character of geometry. We will return to this argument when discussing current approaches to quantum gravity.

Remote as these two chapters may seem from theoretical physics, they serve a double purpose. First, they bring to light a certain argument structure that repeats itself throughout the history of science, and emerges again and again in debates among physicists on the possibility of introducing fundamental length into field theories, and on the consistency of this notion with other well-established physical principles. Second, they emphasize the important question of whether or not the issue of spatial discreteness can be decided empirically. This question is at the heart of the ongoing debate on the quest for quantum gravity phenomenology.

Turning to the history of physics, we begin in Chapter 4 to follow the bumpy road along which the notion of fundamental length entered into field theories. Starting with classical electrodynamics and the problematic idea of the extended electron, via the divergence problems in quantum electrodynamics, and the remedies sought in momentum cutoffs and finite measurement resolution, we learn about the different roles the notion of fundamental length played for different physicists. Moving to quantum field theories we compare the finitist approach (that abhors actual infinity and singularity) with the renormalization program (and its acceptance of potential infinity). Along the way we encounter for the first time the main conceptual problems that accompany fundamental length – postulated at that time to be of the order of the electron radius – namely, non-locality, loss of relativistic causality, and tension with unitarity and Lorentz invariance. These problems (and the strategies for their possible solutions) resurface later in the attempts to "quantize" the general theory of relativity.

Here we also encounter for the first time the different ways in which the fundamental length was interpreted. The high momentum cutoff was initially seen as a limit on spatial resolution, and so the first, natural, interpretation was to regard it as a limit on the domain of applicability of the theory it was introduced to: the theory could not predict anything above the bound on momentum space (or below the bound in position space). Given the numerous problems that one encounters when trying to interpret the high momentum cutoff as signifying an actual spatial discreteness, and the fact that without additional premises there was no logical necessity for doing so, many physicists chose to remain agnostic with respect to the fundamental structure of space. Spatial discreteness thus remained an unsupported conjecture, even after the introduction of a fundamental length into quantum electrodynamics.

As is well known, partly because of the unsolved problems it created in field theories, and partly because of the remarkable experimental and theoretical success of the renormalization program, the notion of fundamental length remained at the margins of theoretical physics. And yet, that quantum field theory is not the arena in which the question of spatial discreteness can be decided follows not from this historical contingency, but from another simple fact, namely that quantum field theory as we know it does not cover all of physics; it leaves out gravitation. We therefore focus our attention on the ongoing attempts to write down a consistent theory of quantum gravity.

In Chapter 5 we lay out – to my knowledge, for the first time – the prehistory of these attempts. Motivated by relationalism and operationalism, mathematicians and physicists, some unknown to contemporary scholars, have entertained ideas germane to the solution of the quantum gravity problem – the problem of constructing a theory that could predict in a consistent way phenomena in domains where both gravitational and quantum field theoretic effects are believed to exist – and did so already in the late 1920s. These ideas include notions such as absolute uncertainty, non-commutative geometry, and non-trivial geometry of momentum space. In this chapter we also discuss the debate that emerged around the late 1950s on the necessity of quantizing the gravitational field, and its putative relation to the notion of fundamental length.

As a preamble to the discussion on current approaches to the quantum gravity problem, we take a short detour in Chapter 6 where we present and analyze a rarely mentioned correspondence between Einstein and the Anglo-American physicist W. F. G. Swann. In this correspondence Einstein reveals his insight into the limit of the "constructive," or "dynamical" approach to spacetime physics, which seeks to reduce spacetime and its geometrical features to relations between more fundamental building blocks and the dynamics thereof. In such an approach, one still needs to designate some degrees of freedom to represent metrical notions such as "length" or "duration," if only to make contact with experiments that can verify the said dynamics in spacetime. Therefore a strong reading of this approach, popular among quantum gravity theorists, according to which "spacetime disappears" at the fundamental level, is untenable. What is tenable is a weaker version of the dynamical approach, which sees the reduction process as a consistency proof. By this I mean a proof for the possibility that the dynamics of some postulated underlying building blocks are consistent with the observable spacetime structure, and may even help decide certain characteristic features of this structure, such as its symmetry groups or its metric. I call this insight "thesis $L$," and I demonstrate its applicability to Swann's unsuccessful attempts to "reduce" (in the strong sense above) the geometrical structure of Minkowski spacetime to the dynamics of quantum field theory.

We further demonstrate the validity of thesis $L$ in Chapter 7, where we show that each of the current contenders for the solution of the quantum gravity problem succumbs to it by designating some degrees of freedom as representing a fundamental notion of "length," "area," or "volume," and does so by stipulation. In addition, each of these approaches either predicts or assumes that the structure of this fundamental geometrical notion is discrete, designating the fundamental length as the Planck length. We further inquire about the role this discreteness plays in these approaches, and find that very little has changed, at least from the conceptual perspective, since the 1930s. First, discreteness is still believed to be the cure for the divergences of quantum field theories, and now, since it includes gravity and hence is imposed almost 20 orders of magnitude below the scale of current observed interactions, it is also believed to be the cure for the singularities of the general theory of relativity. Second, the same difficulties that bothered the physicists who entertained the notion of fundamental length in the 1930s, for example, difficulties with locality, with relativistic causality, with unitarity, and with Lorentz invariance, still haunt us today.

Chapter 8 expands on this last point, focusing on two methodological challenges that quantum gravity theorists face today in their struggle to construct phenomenological models that could turn the question of spatial discreteness into an empirical question. First, it is extremely hard to sacrifice certain well-established principles that support the applicability of the continuous worldview at one – yet to be probed – scale (thus allowing the discrete model to yield new predictions) without violating them at all scales (thus making the discrete model false already). Second, settling for violations *in principle* of these well-established principles, which, given their negligibility or rarity, may never be tested *in practice*, is of no avail if one wishes to turn the question of spatial discreteness into a truly empirical question. In the final part of this chapter we find out whether or not these challenges are currently met.

A concise Q&A summary of the main theses presented and defended in the book can be found in Chapter 9.

Before we start, a disclaimer is in place. Throughout the history of science, many have entertained the notion of fundamental length and interpreted it either ontologically as physical spatial discreteness, or epistemologically as a limit on spatial resolution and on the domain of applicability of the field theory at hand. Owing to restrictions on length (excuse the pun), I have probably left out quite many of these (issues of propriety, for example, were not on my agenda). The historical account I present here mentions most of the major players and a few of the underdogs, yet is by no means exhaustive. Its sole purpose is to harness the history of physics for the sake of extracting philosophical and methodological morals, morals that may be useful to current research in quantum gravity. For "Those who cannot remember the past are condemned to repeat it" ([San05], p. 284).

# 2

# Arguments from math

## 2.1 Outline

The basic mathematical tool, commonly used in the description of physical reality, has long been, and still is, the calculus. This tool is fundamentally continuous: it employs variables that range over continuous sets of values, and the functions it deals with are continuous. One such continuous function describes motion in space, and is treated as a function of a continuous time variable. The continuity of the motion function is essential, for velocity is regarded as the first derivative of this function, and acceleration as the second derivative. Functions which are not continuous are not differentiable, and hence they do not even have derivatives.

Despite its geometrical origins (e.g., [Boy49]), the calculus has been completely "arithmetized," and its development does not require any geometrical concepts. Nevertheless, as a tool, it is still applied to phenomena that occur in physical space. The applicability of the calculus to spatial events is achieved through analytic geometry, which begins with a one-to-one mapping between the points on a line and the set of real numbers. The set of real numbers constitutes a continuum in the strict mathematical (Cantorian) sense; consequently, the order-preserving one-to-one mapping between the real numbers and the points of the geometrical line renders the line a continuum as well. If, moreover, the geometrical line is a correct representation of lines in physical space, then physical space is likewise continuous.

The continuity of the calculus is thus buried deep in standard mathematical physics. Electromagnetism provides a typical example. Here, space is a structure $S$, diffeomorphic to $\mathbb{R}^3$, and the electromagnetic field at each point is an element of $Q = \mathbb{R}^6$, so that the phase space for the whole field is $Q^S = (\mathbb{R}^6)^{\mathbb{R}^3}$. On this phase space, we assign a dynamics in the form of a group of transformations $T$ (time) indexed by $\mathbb{R}$. An impressive infinity for nature to keep track of.

While this is what the common view of physical space amounts to (see, e.g., Salmon in [Sal80], pp. 62–63), many throughout the history of science have felt that the continuum model is an unphysical idealization, and that there are in fact only a finite number of degrees of freedom in any finite volume:

> It always bothers me that, according to the laws as we understand them today, it takes a computing machine an infinite number of logical operations to figure out what goes on in no matter how tiny a region of space and no matter how tiny a region of time. How can all that be going on in that tiny space? Why should it take an infinite amount of logic to figure out what one tiny piece of spacetime is going to do?[1]
>
> *([Fey65], p. 57)*

This tension between the discrete and the continuous has its roots in the philosophy of mathematics, and goes back to the days of Zeno. Indeed, independently of its applicability to physical phenomena, the continuum is one of the most abstract concepts mathematicians have played with, and debates over its cogency have fueled the development of mathematical thought:

> It will always remain a remarkable phenomenon in the history of philosophy, that there was a time, when even mathematicians, who at the same time were philosophers, began to doubt, not of the accuracy of their geometrical propositions so far as they concerned space, but of their objective validity and the applicability of this concept itself, and of all its corollaries, to nature. They showed much concern whether a line in nature might not consist of physical points, and consequently that true space in the object might consist of simple [discrete] parts, while the space which the geometer has in his mind [being continuous] cannot be such.
>
> *([Kan02], p. 288)*

In this chapter we shall take a quick look at these debates, analyzing one of Zeno's arguments against the continuum, as well as other arguments from topology, geometry, and the theory of computation. These *a priori* arguments aim to establish – with varying degrees of sophistication – that only one mathematical structure – be it discrete or continuous – is logically possible, and, furthermore, that only one mathematical structure can be consistently applied as a description of the physical world.

Our goal here, however, will be to defend exactly the opposite claim. We shall try to demonstrate that (1) alternative mathematical descriptions of the line segment are as consistent as the standard, continuous one, and that (2) the question of the applicability of these structures to the physical world cannot be decided by logic alone.

[1] Feynman was introduced to the hypothesis of finite nature in general (and to the notion of cellular automata as a possible simulator to physics in particular) by Edward Fredkin around 1962 (Fredkin, private communication).

## 2.2 Zeno's paradox of extension

### *2.2.1 The paradox*

One of the first people to have inquired about the nature of space (and time) from an *a priori* standpoint was a Greek named Zeno of Elea. A follower of Parmenides, Zeno sought to defend the metaphysical view that denied the possibility of change, motion, and plurality. Consequently, his paradoxes are viewed (e.g., in [Owe01]) as denouncing space, time, and motion as unreal and illusory. Together they form a double attack on *both* the continuous and the discrete.

Zeno's less famous paradox of plurality – also known as the paradox of extension – targets the notion of the continuum. It calls into question the consistency of this notion, and therefore it also underlies the four more famous paradoxes of motion. The latter raise further questions about the nature of motion as a functional relation between the two continua, space and time: Achilles and the Tortoise and the Dichotomy paradoxes target the view that motion is continuous; the Arrow and the Stadium target the view that motion is discrete.

We learn of the paradox of extension a full millennium later than Zeno, from what is supposed to be a direct quotation from him by Simplicus (sixth century AD):

> In his [Zeno's] book, in which many arguments are put forward, he shows in each that a man who says that there is a plurality is stating something self-contradictory. One of these arguments is that in which he shows that, if there is a plurality, things are both large and small, so large as to be infinite in magnitude, so small as to have no magnitude at all. And in this argument he shows that what has neither magnitude nor thickness nor mass does not exist at all. For, he argues, if it were added to something else, it would not increase in size; for a null magnitude is incapable, when added, of yielding an increase in magnitude. And thus it follows that what was added was nothing . . .
>
> The infinity of magnitude he showed previously by the same process of reasoning. For, having first shown that "if what is had not magnitude, it would not exist at all," he proceeds: "But, if it is, then each one must necessarily have some magnitude and thickness and must be at a certain distance from another. And the same reasoning holds good for the one beyond: for it also will have magnitude and there will be a successor to it. It is the same to say it once and to say it always: for no such part will be the last or out of relation to another. So if there is a plurality, they must be both large and small. So small as to have no magnitude, so large as to be infinite."
>
> *([Lee36], pp. 19, 21)*

Zeno's paradox of extension presents the following dilemma: if plurality exists, i.e., if large things are composed of small ones, and infinite divisibility is allowed, then a finitely extended object is supposed to be an (infinite) sum of more basic parts. These basic parts can be extended or unextended. If these parts are unextended, then they have no magnitude, but then their (infinite) sum is also unextended. If these basic parts are extended, then their (infinite) sum is infinite. And so a finite, extended, object is "... both large and small. So small as to have no magnitude, so large as to be infinite."

The issue at stake here is, thus, the convergence (or lack thereof) of an infinite series. It is true that an infinite series of terms may converge to a finite sum when there is no smallest term in the series, but in Zeno's case the ultimate parts are all equal to some $n$ where $n$ is either 0 or some positive number. If the former is the case, the series does converge, but it converges to 0. If the latter is the case, the series diverges and has no finite sum.

### 2.2.2 *Infinite divisibility*

Since the paradox of extension relies on the notion of actual infinite divisibility, one way to resolve it is to deny this notion. Such a denial, however, must overcome the attempts to demonstrate mathematically the logical consistency of this notion that were well known to philosophers already in the seventeenth and eighteenth centuries.[2] These demonstrations fall into two categories: *reductio ad absurdum* proofs (where finite divisibility is assumed and an absurdity is supposed to follow), and proofs by construction. A famous example of the first type of proof is the (in)commensurability of the diagonal of a square with its sides: if both the diagonal and the sides are composed of a certain number of indivisible parts, then one of these indivisible parts would constitute a common measure of the two line segments. Examples of the second type of proof involve constructions that aim to establish infinite divisibility from infinite extendibility, for example, one is asked to imagine a ship that is sailing off in an infinitely extended flat sea. As the ship sails away, its apparent size becomes infinitely small. A combination of these two types, where a construction is used as a basis for a *reductio* is also prevalent in these texts, for example, two concentric circles are used to demonstrate how from the assumption of finite divisibility one derives that the number of points on the circumference of each circle is the same, from which the absurdity that their circumferences are equal in size follows.

We shall not pause to discuss these proofs here, as they were conceived before the development of the modern, set-theoretic, notion of infinity.[3] But in the nineteenth century, with the works of Cauchy, Dedekind, Weierstrass, and Cantor, the notions of limit and convergence became rigorously formalized, and infinities were re-conceived within the axioms of set theory. In this context the line segment is

[2] Most of these demonstrations are found in *The Port-Royal Logic* [Bay51], a logic textbook first published anonymously in 1662, and in Isaac Barrow's *Lectiones Mathematicae*, first published in Latin in 1683.

[3] The above arguments were considered suspect by philosophers such as Hume, Reid, and, more famously, Berkeley, who had already criticized infinitesimals and the shaky foundations of the seventeenth century calculus in *The Analyst* [Fog88, Cum90].