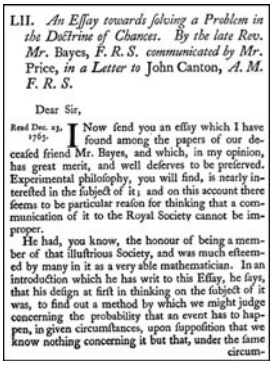
Jahresber Dtsch Math-Ver (2014) 115:129–133
DOI 10.1365/s13291-013-0069-z

HISTORICAL ARTICLE

250 years of "An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F.R.S. communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S."

F. Thomas Bruss

Published online: 18 September 2013



LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763. I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.

He had, you know, the honour of being a member of that illuſtrious Society, and was much eſteemed by many in it as a very able mathematician. In an introduction which he has writ to this Eſſay, he ſays, that his deſign at firſt in thinking on the ſubject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumſtances, upon ſuppoſition that we know nothing concerning it but that, under the ſame circum-

Probably everybody in science has heard the name of (Reverend) Thomas Bayes (born in 1701 or 1702). The Encyclopædia Britannica calls Bayes a "nonconformist theologian (presbyterian) and mathematician". Before his studies in "divinity and logic" he was most likely privately educated. Later he received the major part of his education in Mathematics at the University of Edinburgh. Bayes became minister of his Presbyterian church, first in London and then in Tunbridge Wells. He died in Tunbridge Wells in 1761.

This review is about the work for which he is best known today, that is, his Essay towards Solving a Problem in the Doctrine of Chances, communicated by Price to the Royal Society [1]. It contains among other results a special case of his celebrated inverse-probability formula. Before reviewing his Essay it may be helpful to give some more background.

It was not hard to find out more about the communicator to the Royal Society indicated in the title. It was Mr. Richard Price (1723–1791), a Welsh moral philosopher. Price was active as a writer in radical republican and liberal causes like, among others, the American revolution. Price also wrote on statistics and finance, works less known today, but to which, apparently, he owed his nomination as a Fellow of

This is a slightly revised version of a review which appeared first in Zentralblatt für Mathematik (Zbl. 1250.60007).

F.T. Bruss (✉)
Faculté des Sciences, Dépt. de Mathématique, Université Libre de Bruxelles, CP 210, 1050 Bruxelles, Belgium
e-mail: tbruss@ulb.ac.be

the Royal Society (F.R.S.). The style of his communication advertising Bayes' work shows much respect towards John Canton (a highly regarded physicist), but it also testifies true interest in Bayes' work, and also self-confidence of his own judgment of Bayes' work. Price invested his effort to attract Canton's attention to this work of Bayes. No doubt, much credit should be given to him for this.

The "A.M." part in Canton A.M.F.R.S., by the way, probably just means "Artium Magister" (MA) although this way of aligning a degree and a distinction would nowadays be unusual. I owe this explanation to Frank P. Kelly, F.R.S. Kelly also informed me about S.M. Stigler's paper [2] on which we should comment in the present "classical" review.

First a few comments on the style of Bayes' essay. At the beginning, the reviewer found this essay hard to read. Learning on the way Bayes' terminology made it then easier; nevertheless, although the arguments are rather elementary, this takes still some effort. Here it is amazing to see again what a difference terminology can make. For instance we may wonder nowadays why an author, educated as Bayes was, would not create a word for the binomial coefficient $\binom{n}{k}$ instead of speaking of something like "that coefficient that will be attached to the term containing a^k if the expression $(a+b)^n$ is developed into its parts according ...", but this is the way he *always* wrote.

There are many calculations in this essay. Price was apparently worried that Canton may hesitate for that reason to publish the Essay, pointing out that he does not expect Canton to go through the details. He assured Canton that he himself checked all the calculations and found no mistake, taking all responsibility for possible errors. (How good to know for the present reviewer, who could confine himself to sampled checking, agreeing with Price.)

Now, what exactly is Bayes' famous PROBLEM? It is announced in Sect. I of the Essay and reads:

Given the number of times in which an unknown event has happened and failed: Required the chance that the probability of its (specific event) happening in a single trial lies somewhere between any two degrees of probability that can be named.

The interchange of "chance" and "probability" and terms like "degrees of probability" and others should not confuse us. If we go on and focus on the result we see what is meant. Today we would say: "Given that n independent Bernoulli trials with unknown constant success probability p have brought k successes, what is the probability that the parameter p lies between two given bounds?" And, ironically, we are seduced to ask "Mr. Bayes, do you mean in a Bayesian setting with a given prior density for p?" Indeed, this is what he meant in today's language, and his prior, without saying so, is the uniform prior on [0, 1].

In his essay, Bayes answers this question via several intermediate results he derived, and for which he needed basic definitions. We sample a few of his definitions and notations:

Definition 1.5 The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed and the value of the thing expected upon it's happening.

Indeed, if we get a success with probability p and a reward R upon a success and nothing on a failure it is true that $p = E(\text{reward})/R = Rp/R$. Terms like expecta-

tions or conditional expectations are sometimes imperceptibly interchanged. Again the reading shows he makes no error.

"Inconsistent" means "disjoint", "contrary" means "complementary", etc., but one quickly gets used to this. Sometimes things become little puzzles, however, and not only for non-native English speakers, I guess. See for example Proposition 6 on p. 382: *The probability that several independent events shall all happen is a ratio compounded of the probabilities of each.* Why ratio? Why not the product? What Bayes is bound to have meant is that this probability is the *product* of the respective probabilities which turns out to be a ratio (fraction), of course, since all his probabilities are understood as rational numbers.

As so often, there is reward in going on with the effort of reading. One understands the way Bayes was thinking. The essential point is that he seemingly preferred to think in terms of games and expected rewards, i.e., he had a clear preference to translate probabilities into expectations of gain or loss functions. Interestingly, as much as we profit (nowadays) from the simple identity of the probability of an event A and the expectation of its indicator $\mathbf{1}_A$, the more complicated the arguments become if one does not use this intrinsic link. Bayes had to go through different interpretations of loss functions instead.

Why Bayes thought in terms of expectations of loss function is not clear, in particular as he was essentially concerned with the "uniform" case of a priori equally likely events. Why then not use the intuitive notion as in Laplace's definition of probability of an event A as being the number of cases which are "favourable" for A divided by the number of possible outcomes of experiments?

British education with its traditions may have played here a role. The British are known to like games, and to express probabilistic or statistical statements quite often in terms of games. Perhaps this was always the case. This may explain, by the way, why the word "odds" is not only an English-language creation but can also be heard in Britain at least as often as the word probability. Thinking in terms of games has certainly done no harm to probability theory in Britain and/or to its distinguished probabilists, thus the non-British should not wonder. Also, Bayes may have had independent reasons for his preference.

Thomas Bayes is particularly known for his "formula" to compute inverse conditional probabilities. As Stigler and others point out, Bayes' formula (theorem) is never clearly stated in his essay. He only proved a special form of it, and Proposition 5 is closest to an explicit statement. However, the way to the formula is apparent in several instances of the Essay.

Bayes' Formula As Seen Today. The formula is a versatile tool, both in theory and applications. Clearly, it had an important impact on many problems and it is difficult to imagine it could possibly lose, one day, its interest. Hence it is immortal and one of the jewels in Mathematics one would like to have discovered oneself: very useful, intuitive (I think), and in modern language of Probability very easy to prove. Recall today's notation $(\Omega, \mathcal{F}, P)$ for a probability space and $P(A|B)$ for the conditional probability of an event A given an event B. The latter is defined by $P(A|B) = P(A \cap B)/P(B)$, provided that $A, B \in \mathcal{F}$ and $P(B) \neq 0$. Bayes' "inversion formula" says

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

To prove this equality it suffices to use the definition of a conditional probability and the commutativity of the set operation $\cap$, namely

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \cap A)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}.$$

$P(B)$ can be written for any partition $\{A_1, A_2, \dots\}$ of Ω as $P(B) = \sum_k P(B|A_k) \times P(A_k)$. So we can replace A above by an arbitrarily chosen but fixed A_k. Hence $P(A_k|B)$ is expressed in terms of the $P(B|A_j)$'s and absolute probabilities so that the name "inversion" is adequate.

Let us take an example (which is not in the essay and will only serve as a comparison). We have two urns: urn I contains one black and two white balls, and II contains two black and three white balls. One urn is chosen (in obscurity) according to $P(\mathrm{I}) = 0.4$, $P(\mathrm{II}) = 0.6$, and then one ball is sampled at random from that urn. If the ball is black (B), what is the probability that it stems from urn I? By Bayes' formula we get

$$\begin{aligned} P(\mathrm{I}|B) &= \frac{P(B|\mathrm{I})P(\mathrm{I})}{P(B)} = \frac{P(B|\mathrm{I})P(\mathrm{I})}{P(B|\mathrm{I})P(\mathrm{I}) + P(B|\mathrm{II})P(\mathrm{II})} \\ &= \frac{(1/3)(2/5)}{(1/3)(2/5) + (2/5)(3/5)} = \frac{5}{14}. \end{aligned}$$

A Word on Intuition. The reviewer always liked Bayes' formula because (rare event) he happened to discover it independently in an unprepared exam. Indeed, intuition tells us to look for an equivalent model with a uniform choice of urns and a uniform choice of balls. Here is the solution to the above example in unsophisticated shorthand notation:

$$\left\{ \frac{2}{5} \to (1B, 2W) \;\middle|\; \frac{3}{5} \to (2B, 3W) \right\} \leftrightarrow_{\text{urns}} \big\{ 2(1, 2) \mid 3(2, 3) \big\}$$
$$\leftrightarrow_{\text{balls}} \big\{ 2(5, 10) \mid 3(6, 9) \big\}.$$

The answer is $10/28 = 5/14$ since in the last 5-urn model each ball is equally likely to be chosen, now having 10 black balls on the left and 18 black balls on the right.

What we have done is to create a *uniform urn-and-balls*-model, i.e., adapt the number of urns according to the probabilities $P(\mathrm{I})$ and $P(\mathrm{II})$ and then go over to equal numbers of balls in each urn (respecting the relative frequencies of colours). This works for an arbitrary number of urns and arbitrary contents and may be considered as an algorithmic version of Bayes' formula for rational numbers. This is a very intuitive approach, I think, and for a smaller number of urns and balls by the way not bad as an algorithm. The reviewer found nowhere a reference to this, although he thinks it unlikely that he was the first one seeing this.

This algorithm (or anything similar of this kind) may convince us that many people may have discovered Bayes' Theorem independently. We should keep this in mind for the arguments given below.

Bayes' Formula and Its Origin. Bayes did not yet have the notion of a probability measure, or even "random variable", of course. The latter took roughly one and a half centuries more to be born. Bayes' essay is another instance where we can see that these were major steps in the history of probability, making many things so much easier. Bayes came up with his result through calculations. We are not allowed to be turned off by this; we should appreciate his insight.

We will possibly never know whether Bayes was really the first to see the formula; see Stigler's interesting article [2] for an extensive analysis of this question. Stigler gives an amusing Bayesian argument to indicate that the odds are three to one against him. More precisely, Stigler gave the three for the English mathematician N. Saunderson (1682–1739) who was indeed, for multiple reasons, a very remarkable scientist. The original part of Stigler's argument is that he applied a Bayesian argument to "beat Bayes". We should add that all frequentists of the bad kind would probably enjoy seeing what can be done with priors! The three-to-one result of this analysis of Stigler's is not serious, and not meant to be serious, I guess. Also, with Stigler's "law of eponymy" (cf. [3]), Stigler is allowed to stay faithful to his own dogma and to preach for his chapel, as the French would say. Nevertheless, Stigler's arguments in [2] based on a considerable amount of historical research are relevant and show that certain questions of the true origin of the formula stay open.

This reviewer, possibly biased by his own experience with Bayesian problems reported above, still sees things differently, however. He has little reason to doubt that Bayes' essay is his own and *not* a reproduction of ideas of others. And it is good to see that Stigler is careful to avoid any implication of this kind. Bayes had, as I see it, looked at the typical questions leading to his formula, though the formula itself, we say it again, seems to be nowhere clearly stated. He had good ideas about what he wanted to say on the way to it. If Price claimed that Bayes had solved a problem A. de Moivre could answer only partially, then this can be defended. We should not be confused or even become suspicious by the style of an author who, like Thomas Bayes, had to create his own terminology before being able to express his thoughts.

I also see a true motivation for Bayes' essay. Bayes' research was motivated (this was also true for Price) by religious-philosophical questions, as for instance questions concerning the different proofs of existence of God. For a minister, as Reverend Bayes was in his official position, we agree that such questions must have been a strong motivation to give an answer, and this quite independently of having broader interests.

Taking all these points together (and I believe that Stigler would not disagree) the reviewer can conclude that, unless the contrary is proven, we are all entitled to be faithful to the name Bayes' Theorem.

References

1. Bayes, T.: An essay towards solving a problem in the doctrine of chances. Philos. Trans. R. Soc. Lond. **53**, 370–418 (1763)
2. Stigler, S.M.: Who discovered Bayes' Theorem? Am. Stat. **37**, 290–296 (1983). Zbl 0537.62004
3. Stigler, S.M.: Stigler's law on eponymy. Transactions of the New York Academy of Sciences, Ser. 2, **39**, 147–158 (1980)

Jahresber Dtsch Math-Ver (2014) 115:135–152
DOI 10.1365/s13291-013-0078-y

SURVEY ARTICLE

Punktprozesse in der statistischen Risikoanalyse

Winfried Stute

Online publiziert: 11. Januar 2014

Zusammenfassung Punktprozesse stellen in der Stochastik eine wichtige Klasse von Modellen dar, um zeitliche Abfolgen von „Events“ dynamisch zu erklären. Ziel dieses Artikels ist es, dem Leser auf eine nicht zu technische Art die Bauteile der Modellierung zu beschreiben und Ansätze für ihre statistische Analyse zu präsentieren.

Schlüsselwörter Punktprozesse · Hazardfunktion · Kaplan-Meier Schätzer · Sich Selbsterzeugende Phänomene

Mathematics Subject Classification (2010) MSC 62G55 · 62G30 · 62N01 · 62N02

1 Einführung

Einer verbreiteten (mitunter schlechten) Strategie folgend möchte ich gleich mit einer Definition ins Haus fallen und Ihnen sagen, was man in der Statistik unter einem Punktprozess versteht.

Definition 1 Ein Punktprozess (auf der reellen Achse) ist eine (endliche oder unendliche) strikt monoton wachsende Folge $T_1 < T_2 < \cdots$ von reellen Daten.

Beim Anblick dieser Definition, die eigentlich nur eine Namensgebung ist und keinen in die Tiefe gehenden Sachverhalt festhält, mag man sich fragen, warum die Stochastik dafür einen gesonderten Begriff reserviert hat. Man erinnert sich aus der

W. Stute (✉)
Mathematisches Institut, Justus-Liebig-Universität Gießen, Gießen, Deutschland
e-mail: Winfried.Stute@math.uni-giessen.de

Analysis daran, dass solche Folgen, sofern sie beschränkt sind, konvergieren. In einfachen Übungsaufgaben fordert man dort Studierende auf, spezielle Folgen wie z. B.

$$T_n := \left(1 + \frac{1}{n}\right)^n \tag{1}$$

hinsichtlich ihrer Konvergenz zu untersuchen. Um das Wesen der stochastischen Fragestellungen und Denkweisen besser zu verstehen, möchte ich an (1) anknüpfen, und zwar an einer Stelle, die man zunächst nicht erwartet. Die Notation „:“ beinhaltet insbesondere, dass die Folge mathematisch wohldefiniert ist. Im Rahmen stochastischer Fragestellungen wird sich dies ändern. Der Einfachheit halber beschränken wir uns zunächst auf endliche Folgen, die nur aus T_1 bestehen. Die Wahl der Notation lässt darauf schließen, dass wir uns zum besseren Verständnis unter $T_1 \equiv T$ häufig einen Zeitpunkt vorzustellen haben, zu dem ein uns interessierendes Ereignis (Event) eintritt. Wenn wir $T \geq 0$ voraussetzen, unterstellen wir, dass wir die Zeituhr (des Lebens) zu Beginn auf null setzen. Hier sind fünf Beispiele:

1. $0 \triangleq$ Zeitpunkt einer Krebsoperation und $T \triangleq$ Zeitpunkt der Rückkehr von Metastasen.
2. $0 \triangleq$ Erstimmatrikulation und $T \triangleq$ Zeitpunkt eines erfolgreichen Studienabschlusses.
3. $0 \triangleq$ Eröffnung eines Geschäfts und $T \triangleq$ Zeitpunkt des ersten erfolgreichen Geschäftsabschlusses.
4. $0 \triangleq$ Emission eines Bonds und $T \triangleq$ Defaultzeitpunkt des Emittenten.
5. $0 \triangleq$ Geburt und $T \triangleq$ Zeitpunkt, zu dem uns die „Große Liebe“ unseres Lebens begegnet.

Wir sehen an diesen Beispielen, dass es nicht im mindesten um Konvergenzuntersuchungen geht. Die ganze Geschichte gerät schon früh ins Stocken, da man sich fragen muss, ob es in Einzelfällen überhaupt zur Realisierung von T kommt (keine Metastasen, kein erfolgreicher Studien- oder Geschäftsabschluss). Im fünften Beispiel kommen in der Regel Identifizierungsprobleme (d. h. Blendeffekte) hinzu. Fassen wir zusammen: Im Gegensatz zu Folgen in Analysis oder Numerik begegnen uns in der Stochastik Folgen, die nicht bis ins letzte Detail „wohldefiniert“ sind. Diese Tatsache mag man als Defizit einstufen, trifft aber genau den Kern der Realität, wenn man einmal die interne Welt der Mathematik verlässt. Der Mensch subsumiert derartige Unsicherheiten bzw. Abhängigkeiten von nicht kontrollierbaren Faktoren unter der Überschrift „Zufälligkeit“. Somit werden, um unsere Definition zu ergänzen, die Daten $T_i = T_i(\omega)$ zu Werten, die von auf einem anonymen Zustandsraum Ω definierten Zufallsvariablen angenommen werden und uns mit Informationen über den unbekannten Zustand $\omega \in \Omega$ versorgen. Sehr häufig versieht man Ω mit einem System $\mathcal{A}$ meßbarer Mengen und einem Wahrscheinlichkeitsmaß $\mathbb{P}$, über welche die Verteilungen des Prozesses gesteuert werden. Zeitgleich verwendet man einen zweiten Begriff, der auf die Möglichkeit abrupter Zustandsveränderungen – ohne Vorwarnung – abzielt, den des „Risikos“. Dieser Begriff ist wertneutral zu verstehen, insofern in Anwendungen Veränderungen sowohl zum Guten als auch zum Schlechten denkbar sind.

Bevor wir uns im nächsten Abschnitt mithilfe eines kleinen Beispiels näher mit der Thematik beschäftigen, halten wir fest, was sich hinter Definition 1 eigentlich verbirgt: Ein Punktprozess ist eine zufällige Abfolge von Zeitpunkten, die für die Betroffenen häufig Statusveränderungen bedeuten. Anstelle von Konvergenzuntersuchungen lautet die Hauptfrage, ob es in dieser Welt nicht kausaler Zusammenhänge Mechanismen, Parameter etc. gibt, die dazu beitragen können, die Dynamik dieser Prozesse besser zu verstehen. Sollte es sie geben, dann sicherlich nicht offenkundig, sondern unter Daten verborgen. Denken wir an Anwendungen in der Medizin (Therapien) oder Marktforschung (Werbung), schließen sich weitere Fragen an, z. B. nach der Möglichkeit einer Manipulation (Einflussnahme, nicht Verfälschung). Zudem sollten wir darauf vorbereitet sein, dass Prozesse oder Funktionen, die abrupte Veränderungen modellieren, nicht stetig sind. Sprungzeitpunkte und Sprunghöhen (Gewichte) werden somit zu wichtigen Bausteinen der Statistischen Analyse. Zum Schluss wird es interessant sein zu sehen, wie die Stochastik trotz ihrer Sonderstellung durchaus von Erkenntnissen und Techniken der klassischen (d. h. kausal bzw. deterministisch geprägten) Mathematik profitieren kann.

2 Die Geschichte von Clara und Frederik

Betrachten wir wiederum einen (zufälligen) Zeitpunkt $T = T(\omega)$. Der zugehörige Zählprozess, der sogenannte Single Event Prozess, ist

$$S_t = 1_{\{T \le t\}}, \quad t \ge 0,$$

wobei 1_A die Indikatorvariable zu A ist. Der Graph von S ist einfach, vgl. Abb. 1. Seine wesentlichen Merkmale sind

- Die (nicht strikte) Monotonie
- Der zufällige Sprungzeitpunkt $T = T(\omega)$
- Der Sprung von Null auf Eins in T und Stetigkeit sonst

Als einfachste Treppenfunktion ringt sie Studierenden in der Regel keinen Respekt ab, und meine Begeisterung für Abb. 1 findet keine Nachahmer. Der Grund dafür ist offensichtlich: eine mangelnde Betroffenheit. Nun erwartet man von Studierenden der Mathematik ja vieles, selten aber eine emotionale Anteilnahme. Wir wollen uns anhand einer kleinen Erzählung klarmachen, dass hinter einem T häufig eine „Story" verborgen ist, die uns helfen kann, mehr Verständnis für S_t aufzubringen.

Es ist die Geschichte von Clara und Frederik, zwei Geschwistern, die sich eigentlich immer gut verstanden haben und es deshalb etwas bedauern, dass man sich nicht mehr so häufig sehen kann wie in der Kindheit. So nutzen sie die letzten Sommertage zu einem gemeinsamen Cafe-Besuch. Da Clara noch einige Besorgungen machen

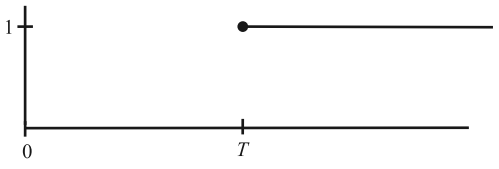


Abb. 1 Graph des Single Event Prozesses

möchte, schlägt sie ihm vor, solange hier zu warten. Es sei jetzt 14:00 Uhr, sie könne nicht genau sagen, wann sie zurück sei, aber um 15:00 Uhr sei sie spätestens wieder da. Frederik, seines Zeichens Mathestudent mit Schwerpunkt Stochastik, schließt daraus, dass die Rückkehrzeit T über dem Intervall [14, 15] gleichverteilt ist und die erwartete Rückkehr bis auf Streuung um 14:30 Uhr erfolgt, eigentlich genug Zeit für ein Bier. Wir drehen die Uhr weiter auf 14:20 Uhr. Clara ist immer noch nicht erschienen, und Frederik passt die erwartete Rückkehrzeit neu an auf 14:40 Uhr. Ähnliches passiert um 14:50 Uhr, wobei ihm unwillkürlich der Gedanke kommt, dass sie ja bald erscheinen müsse. Um kurz vor 14:56 Uhr ist dies dann auch der Fall.

Die Geschichte enthält einige erwähnenswerte Fakten:

1. Zunächst ist F. als Wartender beim Zustandekommen von T unmittelbar betroffen, so wie Sie als Patient oder Investor.
2. Zugleich macht er die Erfahrung, dass er kein Prophet und wie wir alle an das Zeitschema Vergangenheit-Gegenwart-Zukunft gebunden ist.
3. Was ihm bleibt, ist die in der Vergangenheit gesammelten Informationen für eine Prognose zu nutzen und bei Bedarf dynamisch anzupassen.

Mithin ist ein „fertiges“ Bild wie Abb. 1 in der Tat nicht im mindesten dazu geeignet, die Dynamik dieses noch sehr einfachen Prozesses wiederzugeben, wenn es nicht gelingt, die Story „dahinter“ zumindest anzudeuten. Dies herauszufinden sollte Bestandteil einer jeden statistischen Beratung sein.

Die Notwendigkeit, stochastische Prozesse nicht nur nach ihrem Pfadverhalten zu beurteilen, sondern sich zu fragen, welche Information zu welchem Zeitpunkt vorhanden ist und welche Faktoren einen Prozeß antreiben könnten, wurde in der relativ kurzen Geschichte der Stochastik schon früh, d. h. in den 1930er Jahren, erkannt. Zu nennen sind solche Größen wie Kolmogorov, Doob oder Lévy. Welche Folgerungen und Notwendigkeiten sich dabei für statistische Anwendungen (von Punktprozessen) ergeben, soll in den nächsten Abschnitten diskutiert werden.

3 Hazardfunktionen und Kompensatoren

Kehren wir zurück zum Single Event Prozess

$$S_t = 1_{\{T \le t\}}, \quad t \ge 0.$$

Für festes t ist S_t eine 0-1 oder Bernoulli Variable. Die Wahrscheinlichkeit, dass $S_t = 1$, also $T \le t$, bezeichnen wir mit $F(t)$. Die Wahrscheinlichkeit für das komplementäre Ereignis $\{T > t\}$ beträgt

$$\bar{F}(t) = 1 - F(t).$$

$\bar{F}$ heißt häufig Survival Funktion. Sollten Sie den Renditeversprechungen Ihres Bankkundenberaters vertraut haben und Ihr Geld in eine marode Anleihe investiert haben, die zum Zeitpunkt t fällig wird, ist Ihnen zu wünschen, dass die Insolvenz des Emittenten erst zu einem Zeitpunkt T nach t festgestellt wird und Sie ohne Abschläge das Abenteuer überstehen. Die Kenntnis von F bzw. $\bar{F}$ ist also wichtig, um Risiken

richtig beurteilen zu können. Sehr häufig weiß man, wie in unserer kleinen Erzählung, dass T schon die Hürde s geschafft hat. In diesem Fall ist statt F die bedingte Wahrscheinlichkeit

$$\mathbb{P}(T \leq t | T > s), \quad 0 \leq s \leq t, \tag{2}$$

von Interesse. Wir können (2) elementar umschreiben und erhalten

$$\mathbb{P}(T \leq t | T > s) = \frac{\mathbb{P}(s < T \leq t)}{\mathbb{P}(T > s)} = \frac{F(t) - F(s)}{1 - F(s)}.$$

Sollte F differenzierbar mit Dichte (Ableitung) f sein und setzen wir $t = s + \Delta s$ mit $\Delta s \to 0$, ergibt sich

$$\begin{aligned} \mathbb{P}(T \leq s + \Delta s | T > s) &= \frac{F(s + \Delta s) - F(s)}{1 - F(s)} \\ &\sim \frac{f(s)}{1 - F(s)} \Delta s. \end{aligned} \tag{3}$$

Die Funktion

$$\lambda(s) := \frac{f(s)}{1 - F(s)}, \quad 0 < s \tag{4}$$

heißt die zu F gehörige Hazard- oder Intensitätsfunktion. Gleichung (3) besagt, dass die Wahrscheinlichkeit, dass z. B. ein Patient, der nach einer Operation bis zur Zeit s rezidivfrei gewesen ist, bis zum Zeitpunkt $s + \Delta s$ einen Rückfall erleidet, proportional zur Länge Δs des betrachteten Zeitintervalls ist. Ob die Wahrscheinlichkeit groß oder klein ist, hängt somit sehr stark vom Proportionalitätsfaktor $\lambda(s)$ ab. Ein großes $\lambda(s)$ spricht also sehr dafür, dass das Event in Kürze zu erwarten ist. Ist T das reale Sterbealter eines Menschen, spricht man bei λ auch von der Mortalitätsrate.

Falls

$$\bar{F}(t) = 1 - F(t) := \begin{cases} \exp(-\lambda t) & \text{für } t > 0 \\ 1 & \text{für } t \leq 0 \end{cases}$$

die Exponentialverteilung zum Parameter $\lambda > 0$ ist, so wird

$$\lambda(t) = \lambda \quad \text{für } t > 0,$$

d. h. die Hazardfunktion ist gleich dem Parameter und somit konstant. In einführenden Vorlesungen zur Survival Analysis oder Zuverlässigkeitstheorie lernen Studierende viele andere parametrische Familien von Lebensdauerverteilungen kennen. Bezeichnenderweise modelliert man nicht wie sonst die Dichte f, sondern beginnt gleich mit $\lambda(t)$. So gibt es neben dem konstanten λ berühmte monoton fallende und monoton wachsende λ's.

Im Fall von Frederik ist F nach entsprechender Translation gleich der Gleichverteilung $F(t) = t$ auf $0 < t < 1$ mit $f(t) = 1$ dort. Folglich ist $\lambda(t) = \frac{1}{1-t}$ eine Hyperbel, die für $t \uparrow 1$ beliebig groß wird. So ist auch seine Reaktion um 14:50 Uhr zu verstehen, dass sie bald erscheinen müsse.

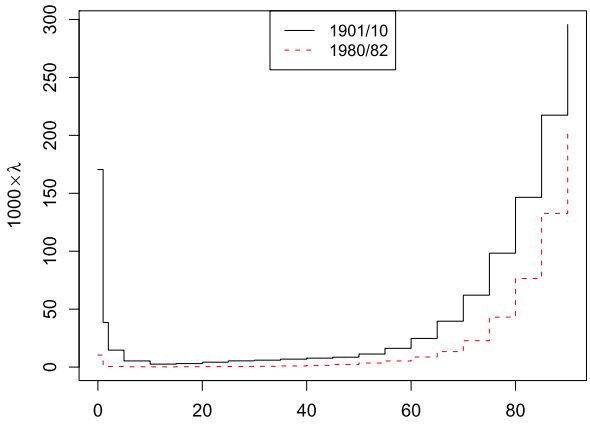


Abb. 2 Hazardfunktionen für die weibliche Bevölkerung in Deutschland in zwei ausgewählten Jahrzehnten

Im nächsten Beispiel beschäftigen wir uns mit der Mortalitätsrate der deutschen Bevölkerung in zwei ausgewählten Jahrzehnten. Abbildung 2 macht deutlich, dass die vom Statistischen Bundesamt veröffentlichten Sterbetabellen zu λ's führen, die eine sogenannte Badewannengestalt haben. Das kurz nach der Geburt erhöhte Risiko („Kindersterblichkeit") ist zwar über Jahrzehnte rapide gesunken, im Vergleich zu etwas älteren Kindern aber immer noch vorhanden.

Die Stammfunktion von λ aus (4) heißt kumulative Hazardfunktion:

$$\Lambda(t) := \int_0^t \frac{f(s)}{1-F(s)}ds = \int_0^t \frac{F(ds)}{1-F(s)}. \tag{5}$$

Wie der Prozess S_t zu Beginn dieses Abschnitts ist auch Λ monoton wachsend. Bei Stetigkeit von f gibt es zwischen Λ und F einen einfachen Zusammenhang:

$$1 - F(t) = \exp\big[-\Lambda(t)\big], \quad t \geq 0. \tag{6}$$

Dabei benutzt man $F(0) = 0$. Mit anderen Worten, Gleichung (5) lässt sich leicht nach F auflösen, und F ist eindeutig durch Λ bestimmt. Schließlich können wir unter Zuhilfenahme der Sprache der Maßtheorie Λ auch als Maß auf $\mathbb{R}^+$ auffassen und erhält

$$d\Lambda = \frac{dF}{1-F}. \tag{7}$$

Wie man leicht sieht, erfüllt die Survival Funktion $\bar{F} = 1 - F$ dann die inhomogene Volterra-Integralgleichung

$$\bar{F}(t) = 1 - \int_0^t \bar{F}(s)\Lambda(ds), \quad t \geq 0. \tag{8}$$

Kommen wir zurück zu (5). Wenn wir im Zähler $1_{\{T \geq s\}}$ hinzufügen, erhalten wir den Prozess

$$t \to \int_0^t \frac{1_{\{T \geq s\}}}{1-F(s)}F(ds) = \int_{[0,t]} \frac{1 - 1_{\{T<s\}}}{1-F(s)}F(ds).$$

Eine letzte Korrektur im Nenner, die die Tatsache berücksichtigt, dass im Zähler die linksseitig stetige Version von (S_t) steht, führt zu

$$A(t) := \int_{[0,t]} \frac{1 - 1_{\{T<s\}}}{1 - F_-(s)} F(ds). \tag{9}$$

Dabei ist $F_-(s)$ der (linksseitige) Grenzwert von $F(x)$ mit $x \uparrow s$. Der Prozess A is selber zufällig und besitzt wie S_t an der Stelle t den Erwartungswert $F(t)$ (Fubini). Der Differenzenprozess

$$M_t = S_t - A_t, \quad t \geq 0 \tag{10}$$

besitzt mithin den Erwartungswert null. Man kann zeigen, dass der Prozess darüber hinaus sogar ein Martingal ist. Dies sind (vereinfachend gesprochen) Prozesse ohne Trend. Der Zähler des Integranden von A verschwindet übrigens für $s > T$, also nach T. Dies ist so gewollt, da in dem von uns bisher studierten einfachen Single Event Prozess mit dem Eintreten des Events die Spannung raus ist. Zur besseren Unterscheidung nennen wir den Integranden von (9) Hazardprozess. In komplexen Punktprozessen muss das Eintreten von T_1 natürlich nicht bedeuten, dass der Hazardprozess bei null verharrt. Stellen wir uns unter den T_i die Zeitpunkte vor, zu denen ein Kunde ein bestimmtes Produkt nachfragt, kann es kurzfristig zu Sättigungseffekten kommen, die sich im Laufe der Zeit wieder auflösen.

Und noch ein letztes: um den stochastischen Indikator in A berechnen zu können, bedarf es lediglich der Kenntnis der Tatsache, ob T im rechts offenen Intervall $[0, t)$ liegt oder nicht. Den Zeitpunkt t muss man nicht abwarten. Man sagt, A ist vorhersehbar. Dies gilt nicht für S_t. Unsere Diskussion hat zum Ziel, uns die eigentliche Rolle von A näherzubringen: wenn $M = S - A$ trendfrei ist und A bereits vor t bekannt ist, böte sich A_t als Prädiktor für S_t an. In A_t ist also das aktuelle Wissen vor t über T hinsichtlich seiner unmittelbaren Zukunft kompensiert. Daher heißt A der Kompensator von S und

$$S = M + A$$

seine Doob-Meyer Zerlegung. Sie ist der Startpunkt für ein Studium von S und anderen Punktprozessen im Rahmen der sogenannten Stochastischen Analysis.

4 Statistik: Nichts als Sorgen

Vielleicht haben Sie mit Genugtuung registriert, dass wir im letzten Abschnitt zu „wohldefinierten“ Begrifflichkeiten zurückgefunden haben. Die Freude wird allerdings nur von kurzer Dauer sein, denn es ist an der Zeit, eines der wichtigsten Wörter im Sprachschatz der Statistik einzuführen: „Leider“. Als Zwischenfazit müssen wir nämlich konstatieren, dass F und Λ leider nicht bekannt sind. Unglücklicherweise trifft auf die Dichte f und die Hazardfunktion λ nicht zu, dass sie uns „gegeben“ sind, wie es häufig so schön in den Theoremen der Mathematik heißt. Auch das Statistische Bundesamt wäre nicht in der Lage, Sterbetafeln zu erstellen, hätte es nicht die Informationen aus den örtlichen Rathäusern. Diese Informationen werden immer bruchstückhaft sein, aber umso kompletter, je mehr Daten eingehen. Den

Vorgang der Informationsverarbeitung nennt man „Schätzen“. Davorgeschaltet ist die Datengewinnung, sie kostet in der Regel viel Zeit und Geld. Fassen wir zusammen: sämtliche bisher vorgestellten mathematischen Objekte sind und bleiben in der Regel leider unbekannt. Datengewinnung und -verarbeitung haben das Ziel, diese unbekannten Objekte zu approximieren. Statistische Untersuchungen setzen sich zum Ziel, die Qualität der Anpassung zu quantifizieren und etwaige Modellannahmen zu überprüfen (testen).

Wir setzen uns im Folgenden zum Ziel, die einzelnen Schritte für Punktprozesse zumindest anzudiskutieren. Die erste wichtige Beobachtung ist, dass die Sprache der abstrakten Maß- und Integrationstheorie, auch wenn die Objekte tatsächlich unbekannt sein sollten, von großem Nutzen sein kann. Gleichzeitig wird deutlich, dass das Lebesgue-Maß seine dominierende Rolle eingebüßt hat und oft nur als Referenzmaß von Dichten (siehe (5)) dient. Auf der anderen Seite tritt ein Maß in den Vordergrund, das versteckt bereits in Abb. 1 enthalten ist, das Dirac- oder Punktmaß. Dazu sei T irgendein Element irgendeiner Menge. Das Dirac-Maß zu T ist dann gegeben durch

$$\delta_T(A) = \begin{cases} 1 & \text{falls } T \in A \\ 0 & \text{sonst.} \end{cases}$$

Dem Punkt T, d. h. der einelementigen Menge $\{T\}$, wird somit die Gesamtmasse eins gegeben, und in Abschnitt 2 ist S_t gerade die Verteilungsfunktion von δ_T. Ist wie in der Stochastik $T = T(\omega)$ zufällig, so auch δ_T. Die Regeln der Maß- und Integrationstheorie gelten natürlich auch weiterhin. Es wird sich herausstellen, dass Dirac-Maße für uns die elementaren Bausteine eines ganzen Gebäudes darstellen.

In der ersten Ausbaustufe halten wir fest, dass ein Dirac-Integral einer Funktion φ gleich dem Wert an T ist:

$$\int \varphi d\delta_T = \varphi(T). \tag{11}$$

Hat man nun statt einem T Daten $T_1, \ldots, T_n$, wobei n der Stichprobenumfang sei, können wir für jedes T_i das Dirac-Maß bilden und die Gesamtmasse eins gleichverteilen. Dies führt zum empirischen Maß

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{T_i}$$

mit zugehöriger empirischer Verteilungsfunktion

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{T_i \le t\}}.$$

S_t entspricht $F_n(t)$ im Fall $n = 1$. Das μ_n-Integral von φ wird wegen (11) zu

$$\int \varphi dF_n = \frac{1}{n} \sum_{i=1}^{n} \varphi(T_i). \tag{12}$$

Die Größen (12) heißen empirische Integrale und sind Schätzer der unbekannten Größe $\int \varphi dF$. Das Ersetzen des unbekannten F durch den Schätzer F_n nennt man plug-in Technik. Für $\varphi(x) = x$ liefert (12) das arithmetische Mittel als Schätzer für die mittlere Lebensdauer.

Wenn wir für φ Indikatoren $1_{(-\infty,t]}$ wählen, kann es sein, dass T_i auf den Rand fallen, d. h. im Gegensatz zum Lebesgue-Maß wird es nun wichtig, zwischen Integralen $\int_{(-\infty,t)} dF_n$ und $\int_{(-\infty,t]} dF_n$ zu unterscheiden. Dies bedeutet im Endeffekt, dass Schreibweisen wie (5) nicht mehr akzeptiert werden können.

Wie bereits in (7) angedeutet, setzen wir für jede Lebensdauerverteilung G

$$d\Lambda = \frac{dG}{1 - G_-}$$

und erhalten statt (8)

$$\bar{G}(t) = 1 - G(t) = 1 - \int_{[0,t]} (1 - G_-)(x)\Lambda(dx). \tag{13}$$

Die Einführung von G erfolgt, weil damit deutlich wird, dass das unbekannte interessierende F sich im „Ozean“ aller G's versteckt hält und durch $G = F_n$ approximiert wird. Da wir das „wahre“

$$\Lambda(t) = \int_{[0,t]} \frac{dF}{1 - F_-}$$

leider nicht kennen, wird es nach dem plug-in Verfahren durch

$$\Lambda_n(t) = \int_{[0,t]} \frac{dF_n}{1 - F_{n-}}$$

geschätzt. Abbildung 3 zeigt, dass die Sprungstellen dieses Prozesses an den Daten liegen, aber nicht mehr gleich groß sind. Λ_n heißt in der Literatur Aalen-Nelson Schätzer. Die T_i, die Abb. 3 zugrundeliegen, gehören zu einer geordneten Stichprobe von $n = 100$ unabhängig und nach einer Exponentialverteilung (mit $\lambda = 1$) simulierten Größen. Mithin ist $\Lambda(t) = t$. Der Schätzer Λ_n ist bis zu $t = 3$ hervorragend, aber in den rechten Flanken nicht akzeptabel. Dies liegt daran, dass Λ unbeschränkt ist für $t \uparrow \infty$ während Λ_n ab dem größten Datum konstant ist.

Sind die T_i durch Anordnung von n unabhängigen identisch nach F verteilten Zufallsvariablen $X_1, \ldots, X_n$ entstanden, ist der Kompensator von F_n gegeben durch

$$t \to \int_{[0,t]} \frac{1 - F_{n-}(s)}{1 - F_-(s)} F(ds).$$

Der Zähler des Hazardprozesses verringert sich zu jedem T_i um $1/n$, was die Tatsache widerspiegelt, dass mit wachsendem s ein T_i nach dem anderen „abgearbeitet“ wird.

Unser nächster Kommentar betrifft (13). Wir können das Integral auf der rechten Seite bei gegebenem Λ als Operator in $1 - G_-$ deuten. Da wir nicht nur differenzierbare Λ, sondern auch solche (wie Λ_n) mit Sprunganteilen haben, greift (6) nicht.

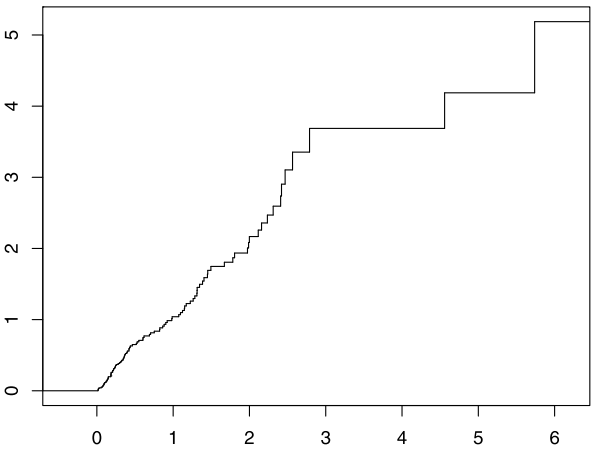


Abb. 3 Aalen-Nelson Schätzer für $n = 100$

Die Inversion von (13) erfordert eine Technik, die uns zum sogenannten Produkt-Limes Integral führt.

Sei dazu Λ eine nichtnegative monoton wachsende rechtsseitig stetige Funktion mit möglichen Sprüngen. Bezeichnen wir mit Λ^c und Λ^d den stetigen bzw. Sprunganteil von Λ, so gilt für die Lösung von (13)

$$1 - G(t) = \exp\bigl[-\Lambda^c(t)\bigr] \prod_{s \le t} \bigl[1 - \Lambda^d\{s\}\bigr]. \tag{14}$$

Hierbei ist $\Lambda^d\{s\}$ die Sprunghöhe von Λ bzw. Λ^d in s. Ist Λ stetig, also $\Lambda = \Lambda^c$, wird (14) zu (6). Dagegen ergibt sich

$$1 - G(t) = \prod_{s \le t} \bigl[1 - \Lambda\{s\}\bigr] \tag{15}$$

falls Λ von reinem Sprungtyp ist und somit $\Lambda = \Lambda^d$ und $\Lambda^c = 0$ gilt.

Die ganze Bedeutung von (14) wird deutlich, wenn wir uns klarmachen, dass es in der Praxis nicht nur Funktionen wie F und Λ sind, die leider unbekannt sind, sondern mitunter Daten selber. Eine Konsequenz wäre, dass uns eine solch banale Aufgabe wie die Berechnung des arithmetischen Mittels vor unlösbare Probleme stellt, weil einige Summanden schlichtweg fehlen. Stellen wir uns dazu im Rahmen einer medizinischen Studie vor, dass man bei n Patienten an den individuellen Zeitdauern X_i bis zum Auftreten eines Rezidivs interessiert ist. In der Regel ist die Projektphase zeitlich befristet, so dass es Patienten gibt, bei denen anstelle der rezidivfreien Zeit lediglich die Aufenthaltsdauer Y_i in der Studie bekannt ist. Mathematisch führt dies zu sogenannten Rechtszensierungen, d. h. statt X_i sind nur die Größen

$$Z_i = \min(X_i, Y_i) \quad \text{und} \quad \delta_i = 1_{\{X_i \le Y_i\}}$$

bekannt. Dabei gibt δ_i den Status des Datums (zensiert oder nicht-zensiert) an. Der gleichen Problematik begegnet man, wenn Studiendauern analysiert werden sollen

und man sich fragt, wie man die Zeitdauern der Studierenden gewichten soll, die ihr Studium noch nicht abgeschlossen haben.

Es stellt sich heraus, dass man in derlei Situationen zur Schätzung von $\bar{F}$ einen Umweg über Λ zu gehen hat. Dies führt zu einem von Kaplan und Meier [10] vorgeschlagenen Schätzer, bei dem die Gewichte der größenmäßig geordneten Z_i, sagen wir wieder $T_1 < \cdots < T_n$, nun nicht mehr $1/n$ sind, sondern

$$W_{in} = \frac{\delta_{[i:n]}}{n-i+1} \prod_{k=1}^{i-1} \left[1 - \frac{\delta_{[k:n]}}{n-k+1} \right].$$

Dabei ist $\delta_{[i:n]}$ der Status von T_i. Wir sehen, dass die Gewichte von den Labels abhängen und somit zufällig sind. Ein Datum, welches zensiert ist ($\delta_{[i:n]} = 0$) erhält das Gewicht null. Der Schätzer der sogenannten mittleren Disease Free Survival Time wäre nun $\sum_{i=1}^{n} W_{in} T_i$ und hätte nicht mehr die einfache Form eines arithmetischen Mittels. Wenn alle δ_i eins sind, wird W_{in} wieder zu $1/n$. Ist T_n zensiert, geht Masse verloren und die Summe der W_{in} ist echt kleiner als eins. Wir werden in Abschnitt 6 auf diesen Umstand zurückkommen.

Die Herleitung der W_{in} macht eine wesentliche Eigenschaft von $d\Lambda$ deutlich, die nicht von dF geteilt wird. Bezeichnen wir mit G die (leider unbekannte) Verteilungsfunktion der Y_i und ist jedes Y_i von seinem X_i unabhängig, so besitzt die Verteilungsfunktion H der Minima Z_i die Eigenschaft

$$\bar{H} = \bar{F}\bar{G}.$$

Ferner erfüllt die Subverteilung

$$H^1(t) = \mathbb{P}(Z \le t, \delta = 1) = \mathbb{P}(X \le t, X \le Y)$$

die Beziehung

$$H^1(t) = \int_{[0,t]} \big[1 - G_-(s)\big] F(ds)$$

bzw.

$$dH^1 = (1 - G_-) dF.$$

Es folgt

$$d\Lambda = \frac{(1-G_-)dF}{\bar{H}_-} = \frac{dH^1}{\bar{H}_-}.$$

H und H^1 sind aber Verteilung und Subverteilung der beobachtbaren Z und (Z, δ) und somit empirisch mit $1/n$ Gewichten schätzbar. Wenn wir die entsprechenden $\bar{H}_n$ und H_n^1 einsetzen, führt dies zu einem Λ_n (unter Zensierung) und schließlich mithilfe von (15) angewandt auf Λ_n zum Kaplan-Meier Schätzer. Die Arbeit von Kaplan und Meier [10] ist die meist zitierte Arbeit der gesamten Statistik. Vgl. Ryan und Woodall [15]. Dies liegt zum größten Teil daran, dass Lebensdaueranalysen Teil zahlreicher medizinischer Veröffentlichungen sind. Eine Übersicht über die Geschichte der Kaplan-Meier Gewichte findet sich in Stute [19].

Sollten Sie sich bereits auf die nächste Hiobsbotschaft eingerichtet haben, ist das ganz in meinem Sinne. Wie wir bereits gesehen haben, ist man in der Praxis häufig gezwungen, mit zensierten Daten vorlieb zu nehmen. Die Verfügbarkeit der Z_i setzt allerdings voraus, dass Patienten im Rahmen einer Nachsorge ständig kontrolliert werden. Dies ist aus organisatorischen und Kostengründen jedoch nicht möglich, so dass man sich darauf einigt, Kontrolluntersuchungen in einem bestimmten Zeitgitter durchzuführen. In letzter Konsequenz sind damit auch die Z_i unbekannt. Stattdessen ist zu jedem Zeitpunkt nur bekannt, ob zwischen der letzten und der aktuellen Kontrolle ein Rückfall eingetreten ist oder nicht. Diese Situation nennt man in der Literatur „Current Status Model". Darüber hinaus gibt es zahlreiche Varianten, die von der Doppelt- (d. h. beidseitigen) Zensierung bis zum sogenannten „Sacrificing" in der pharmazeutischen Forschung reichen. In AIDS-Studien ist man mit dem Problem konfrontiert, dass die Zeiten $T_1 =$ Infektion und $T_2 =$ Seroconversion unter Umständen nicht bekannt sind und der Fall als Dunkelziffer keine Aufnahme in eine Studie findet. Das Beispiel des Kaplan-Meier Schätzers zeigt, dass es darauf ankommt, statt mit $1/n$ Daten effizient so neu zu gewichten, dass aufgetretene Verzerrungen nicht ausgeschlossen, aber so gut wie möglich reduziert werden. Ideen zur Neugewichtung erhält man aus entsprechenden Operatorgleichungen, die von Fall zu Fall unterschiedlich sind und natürlich, bevor es zur Lösung kommt, erst einmal gefunden werden müssen.

Allgemein gesprochen kann es passieren, dass die Datenlage so schlecht ist, dass wir es im Sinne von Hadamard mit schlecht gestellten inversen Problemen zu tun bekommen. Insgesamt verdichtet sich der Eindruck, dass gute statistische Praxis ein genaues Hinschauen erfordert und keinen Automatismus darstellt. Im Umkehrschluss kommen auf den mathematisch ausgerichteten Statistiker immer dann neue Herausforderungen zu, wenn bisherige Modelle komplexen Datenstrukturen nicht gerecht werden. Einige dieser Aspekte wollen wir im nächsten Abschnitt diskutieren.

5 Individualisierte Statistik

Rufen wir uns noch einmal in Erinnerung, welche Punktprozesse bisher diskutiert worden sind. Ausgangspunkt waren unabhängige identisch verteilte Zufallsvariable X_i, die nach einer unbekannten Verteilungsfunktion F verteilt seien. Die monotone Folge $T_1 < T_2 < \cdots < T_n$ erhält man durch Ordnen der X_i. Stellen wir uns unter den X_i wiederum Lebensdauern von Patienten vor, so muss man sich die Frage stellen, ob die Annahme der identischen Verteiltheit wirklich gerechtfertigt ist, oder ob es nicht andere Faktoren wie Geschlecht, Alter, Raucherstatus etc. gibt, die einen Einfluss auf die Lebensdauer haben und in denen sich Patienten unterscheiden. In diesem Fall würden sich auch die λ unterscheiden und von individuellen Faktoren abhängen.

Ein klassisches und sehr populäres Modell geht auf den englischen Statistiker Cox [3] zurück. Es wird unterstellt, dass λ_i von der Form

$$\lambda_i(t) = \lambda_0(t) \exp\left[\sum_{j=1}^{k} \beta_j F_{ij}\right], \quad 1 \le i \le n, \tag{16}$$

ist. Dabei ist λ_0 (eine leider unbekannte) Hazardfunktion, die allen Personen gemeinsam ist (die sogenannte Baseline Funktion), F_{ij} sind die relevanten Faktoren von Person Nummer i zu Beginn der Studie und β_j ist der (unbekannte) Hebel, mit dem der j-te Faktor in das Risikoprofil des i-ten Patienten eingeht. Das erstrangige Ziel besteht (häufig) darin, die unbekannten λ_0 und β_j zu schätzen. Eine zweite Aufgabe wäre es, die Qualität der Anpassung kritisch zu überprüfen. Zum Beispiel unterstellt (16), dass der individualisierte exp-Teil lediglich von Faktoren abhängt, wie sie sich zu Beginn der Studie darstellen.

Vergleichen wir die Patienten i und m zum Zeitpunkt t hinsichtlich ihres Risikos, ergibt sich

$$\frac{\lambda_i(t)}{\lambda_m(t)} = \frac{\exp[\sum_{j=1}^{k} \beta_j F_{ij}]}{\exp[\sum_{j=1}^{k} \beta_j F_{mj}]}. \tag{17}$$

Die rechte Seite hängt nicht mehr von t ab. Sollte ihr Zähler größer sein als der Nenner, also Patient i im Vergleich zu m bei Feststellung der Risikofaktoren zu Beginn der Studie ein größeres Risiko haben, so unterstellen (16) und (17), dass i auch mittel- und langfristig weiter benachteiligt wird und eine Umkehrung der Risiken ausgeschlossen ist. Gleichung (17) heißt in der Literatur „Proportional Hazards"-Eigenschaft.

Im Folgenden wollen wir Situationen diskutieren, bei denen sich ein Punktprozess nicht allein durch Anordnung (Überlagerung) von (individualisierten oder nicht-individualisierten) Single Event Prozessen erzeugen lässt. Betrachten wir dazu ein Beispiel, bei dem Sie selbst betroffen sind. Wir bezeichnen mit $T_1 < T_2 < T_3 < \cdots$ die Zeitpunkte im nächsten Monat, zu denen Sie einen Einkauf tätigen. Die Anzahl $N(t)$ dieser Zeitpunkte vor t ist von Kunde zu Kunde unterschiedlich, und Kaufsummen (sogenannte Markierungen) wollen wir auch ausblenden. Wie bisher soll $\lambda(t)$ die Bereitschaft messen, dass kurz nach t ein Kauf getätigt wird. Natürlich wird bei dieser Kurznotation unterdrückt, dass λ wichtige Größen wie persönliches Einkommen, individuelle Vorlieben, saisonale Komponenten oder Sättigungseffekte enthält. Diese können sich im Verlauf der Zeit individuell ändern, so dass die Annahme von Proportional Hazards kaum aufrecht zu erhalten ist.

Kennen Sie Ihr persönliches λ? Wenn nicht, ist das nicht weiter tragisch. Es gibt andere, die sich Ihrer annehmen, individualisiert versteht sich, und auf das, was sich in Ihrem Unterbewusstsein abspielt, Einfluss nehmen wollen. Ich darf Sie also ermuntern, zwecks besserer Durchleuchtung möglichst viele (elektronische) Spuren zu hinterlassen. Sollten Sie nicht zufriedenstellend funktionieren oder gar wegen Überlastung zum Vergessen neigen, wird der Druck erhöht. Abbildung 4 ist einer Arbeit von Kopperschmidt und Stute [13] entnommen, wo einerseits die mathematisch statistische Theorie für solche sich selbsterzeugende Phänomene bereitgestellt wird und andererseits Modelle für das Kaufverhalten von Konsumenten entwickelt und statistisch ausgewertet werden. Die Daten basieren auf mehr als 2000 Punktprozessen, die uns über eine international agierende Marktforschungsgesellschaft von den beteiligten Haushalten über mehr als 18 Monate geliefert wurden und Informationen über Kaufakte und TV-Werbekontakte enthielten. Ziel einer Werbemaßnahme ist natürlich, Einfluss auf Ihr individualisiertes λ zu nehmen und es nach oben zu treiben. Bei der Modellbildung des Hazardprozesses λ, die in Absprache mit Experten aus

der Praxis geschah, musste man davon ausgehen, dass die Konsumneigung saisonalen Schwankungen unterlag. Die allen Haushalten gemeinsame Baseline-Funktion wurde

$$\lambda_1(t) = \alpha \sin(\beta t + \gamma) + \delta$$

gesetzt. Der unbekannte Parameter δ repräsentiert die Grundrate, mit der unabhängig von der Jahreszeit das Produkt nachgefragt wird. Der zweite (zufällige) Baustein besteht aus einem indvidualisierten Teil, der die interne Kaufgeschichte von Haushalt i festhält, insbesondere die Zeitspanne zwischen t und dem Zeitpunkt $Y_{iN_i(t-)}$ des letzten Kaufakts vor t. Dies führt zu

$$\lambda_2^i(t) = \left(1 - e^{-\varepsilon(t - Y_{iN_i(t-)})} 1_{\{t > Y_{i1}\}}\right).$$

Der (positive) Parameter ε gibt an, wie schnell Sättigungseffekte sich auflösen. Der letzte Baustein enthält Größen, die einerseits den Effekt ξ einer Werbemaßnahme und zum zweiten sogenannte Adstock (Vergessen-)Phänomene in Form eines (negativen) Parameters η einbauen:

$$\lambda_3^i(t) = \xi \sum_{h=1}^{W_i(t)} e^{\eta(t - X_{ih})}.$$

$W_i(t)$ bezeichnet die Anzahl der bis t gemachten TV-Werbekontakte von Haushalt i, und X_{ih} sind deren Zeitpunkte. Zu jedem X_{ih} springt λ_3^i um ξ, um dann mit einer exponentiellen Rate wieder abzufallen. Derlei zufällige Effekte nennt man Shot Noise.

Das endgültige λ wurde

$$\lambda_{\vartheta,i} = \lambda_1 \lambda_2^i + \lambda_3^i$$

gesetzt. Der Parameter ϑ setzt sich aus allen Einzelparametern zusammen, ist also 7-dimensional. Die involvierten Kauf- und Werbekontaktprozesse N_i und W_i werden nicht spezifiziert und sind somit nichtparametrischer Natur. Insgesamt handelt es sich bei $\lambda_{\vartheta,i}$ um ein semiparametrisches Modell. Da in $\lambda_{\vartheta,i}$ interne und externe Einflüsse aus der Vergangenheit vermischt sind, nennt man diese Punktprozesse sich selbsterzeugend. Zusätzlich wurden noch soziodemographische Schichtungen vorgenommen, und zwar nach Einkommen und Größe des Haushalts. Das Bild, welches entsteht, zeigt eine Abfolge von „stetigen“ Phasen und Sprüngen. Sind diese positiv, waren dies Reaktionen auf einen Werbekontakt, während negative Sprünge auf kurzfristige Sättigungeffekte hindeuten.

Was die Schätzung unbekannter Parameter anbetrifft, darf ich an unsere kurze Diskussion in Abschnitt 3 erinnern. Anstelle von (10) hat man nun die Differenz zwischen den beobachtbaren Punktprozessen und den modellierten Kompensatoren zu betrachten und deren Abstand zu minimieren, d. h. Prädiktoren und tatsächliche Daten bestmöglich in Einklang zu bringen. Wenn man will, handelt es sich dabei um eine Weiterentwicklung des Kleinst-Quadrate Prinzips (in der Regression) auf dynamische Punktprozesse.

Insgesamt sollte ein Bild von einer Realität entstanden sein, wie bei einem Seismogramm, in der der Mensch (hier als Kunde) dem Spannungsfeld interner und externer

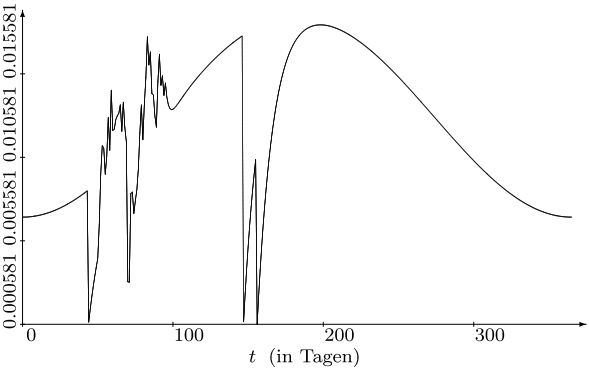


Abb. 4 Individualisierte Kaufintensität eines Haushalts

Wünsche und Kräfte ausgesetzt ist. Man beachte, dass an keiner Stelle von einer in vielen statistischen Verfahren gemachten Annahme einer Stationarität Gebrauch gemacht wird, sondern allgemeine Dynamiken zugelassen sind. Dies liegt daran, dass in medizinischen Studien oder in der Marktforschung auf Punktprozesse von mehreren Patienten oder Kunden zurückgegriffen werden kann. Zusammen stellen sie so viel Information bereit, dass auf einschneidende Annahmen wie Stationarität verzichtet werden kann. Für eine Übersicht von anderen populären Kaufverhaltensmodellen sei auf Kopperschmidt und Stute [12] verwiesen. Erste Modellierungen von sich selbsterzeugenden Punktprozessen in einem anderen Zusammenhang finden sich bei Hawkes [8].

6 Das Drama der Multivariaten Survival Analysis

Die Multivariate Survival Analysis beschäftigt sich mit den Ausfallzeiten von Systemen, die aus mehreren Komponenten bestehen. Im einfachsten Fall ist $X_i = (X_{i1}, X_{i2})$ ein zufälliger zweidimensionaler Vektor. Jede einzelne Koordinate entspricht also einem X_i von früher. Von Interesse ist nun die sogenannte gemeinsame Verteilung

$$F(t_1, t_2) = \mathbb{P}(X_{i1} \leq t_1, X_{i2} \leq t_2).$$

In der Regel sind X_{i1} und X_{i2} z. B. demselben Patienten zuzuordnen, so dass davon auszugehen ist, dass X_{i1} und X_{i2} voneinander abhängen. Sind alle X_i beobachtbar, ist F nichtparametrisch durch die empirische Verteilungsfunktion

$$F_n(t_1, t_2) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_{i1} \leq t_1, X_{i2} \leq t_2\}}$$

schätzbar. Unterliegen wie im eindimensionalen Fall die X_i jedoch dem Risiko, durch einen nun zweidimensionalen Vektor zensiert zu werden, stellt sich die Frage, wie das Analogon zum Kaplan-Meier Schätzer auszusehen hat.

Erste wichtige Beiträge gehen auf Dabrowska [4] zurück, die über die Einbeziehung von bedingten Verteilungen den bivariaten Fall auf den eindimensionalen Fall

reduzierte. Dieser Zugang führt jedoch zu einem schlecht gestellten Problem, welches eine Regularisierung erforderlich macht und letzten Endes Schätzer produziert, die nicht monoton sein müssen. In der Folgezeit gelingt es Autoren wie Gill et al. [7], die Produkt-Limes Integration auf den mehrdimensionalen Fall zu übertragen. Der letzte Schritt zur effizienten Schätzung von F unter Zensierung, also die Erweiterung des Kaplan-Meier Schätzers auf den multivariaten Fall, findet allerdings nicht statt. Der Grund für das Nichtgelingen kommt einem Desaster gleich. Es stellt sich heraus, dass im multivariaten Fall eine Verteilungsfunktion nicht mehr durch die Hazardfunktion eindeutig bestimmt ist.

Damit entfällt die Möglichkeit, eine Survival Funktion im Mehrdimensionalen über die Hazardfunktion zu identifizieren und zu schätzen. Ein erfolgreicherer Zugang nutzt die Tatsache aus, dass bereits beim klassischen Kaplan-Meier Schätzer die Gesamtgewichte kleiner als eins sind, wenn das größte Datum zensiert ist. Die verlorengegangene Masse können wir uns auf den Punkt $t_\infty = \infty$ verschoben denken. Mithin macht es Sinn, bei der Suche nach effizienten Schätzern (im Zweidimensionalen) auch solche Masseverteilungen ins Kalkül mit aufzunehmen, die (kleine) positive Masse auf $t_\infty = (\infty, \infty)$ haben. Bezeichnen wir die Survival Funktion von $X = (X_1, X_2)$ mit

$$\bar{F}(t_1, t_2) = \mathbb{P}(X_1 \geq t_1, X_2 \geq t_2),$$

so besitzt der Zufallsvektor, der aus X entsteht, indem man eine vorgegebene positive Masse ε nach t_∞ schiebt, das Hazard-Maß

$$d\Lambda_\varepsilon = \begin{cases} \frac{(1-\varepsilon)dF}{(1-\varepsilon)\bar{F}+\varepsilon} & \text{auf } \mathbb{R}^2 \\ 1 & \text{auf } t = t_\infty \end{cases}$$

und die Survival Funktion

$$\bar{F}_\varepsilon = \begin{cases} (1-\varepsilon)\bar{F} + \varepsilon & \text{auf } \mathbb{R}^2 \\ \varepsilon & \text{auf } t = t_\infty. \end{cases}$$

Ferner erfüllt $\bar{F}_\varepsilon$ die homogene Volterrra Integralgleichung

$$q_\varepsilon(x) = \int 1_{\{t \geq x\}} q_\varepsilon(t)\Lambda_\varepsilon(dt). \tag{18}$$

Dabei hat die Integration über $\mathbb{R}^2 \cup \{t_\infty\}$ zu erfolgen, und $\geq$ ist koordinatenweise zu verstehen. Nun ist aus der Funktionalanalysis bekannt, dass jede homogene Volterra Gleichung nur die triviale Lösung null besitzt. Bei einer möglichen Übertragung auf unseren Fall hat man jedoch Vorsicht walten zu lassen, da die Resultate für das Lebesgue Maß formuliert werden. Unsere Masseverschiebung hat aber gerade das Ziel, Punktmassen zu erzeugen, die (18) zwar optisch als homogen erscheinen lassen, faktisch aber Inhomogenitäten erzwingen. Man beachte ferner, dass $d\Lambda_\varepsilon$ für $\varepsilon > 0$ ein finites Maß ist, während das ursprüngliche $d\Lambda$ in der Regel infinit ist. Was die Identifizierbarkeit anbetrifft, lässt sich zeigen, dass die Lösung von (18), welche nichtnegativ ist mit $q_\varepsilon(0,0) = 1$ und $q_\varepsilon(t_\infty) = \varepsilon$, eindeutig ist. Diese Aussage gilt nicht nur für $d\Lambda_\varepsilon$, sondern für jedes andere finite Maß P, insbesondere auch für Schätzer $d\hat{\Lambda}_\varepsilon$

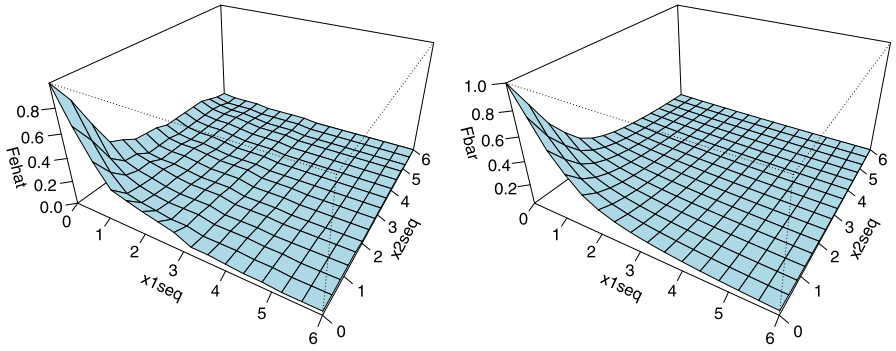


Abb. 5 Bivariate Survival Funktion (*rechts*) und KM-Schätzer (*links*)

von $d\Lambda_\varepsilon$. Die gewünschte Erweiterung des Kaplan-Meier Schätzers erhält man als Lösung von (18) mit $d\hat{\Lambda}_\varepsilon$ anstelle von $d\Lambda_\varepsilon$. Für die Effizienz hat man $\varepsilon \sim 1/n$ zu setzen, wobei n wiederum der Stichprobenumfang ist. Außerdem erhält man für die Lösung von (18) eine interessante Darstellung in Form einer Neumann Reihe, die für Verteilungsaussagen des Schätzers von Nutzen ist. Für die numerische Lösung von (18) wurde ein Algorithmus entwickelt, der den Schätzer in endlich vielen Schritten berechnet. Details finden sich in Sen und Stute [16]. Zur Illustration fügen wir zwei Plots an, rechts die der Simulation zugrundegelegte Survival Funktion und links der oben beschriebene Schätzer (Abb. 5). Der Stichprobenumfang war $n = 100$, und der Prozentsatz der zensierten Daten war 20 %.

7 Zur Literatur

Es gibt mittlerweile zahlreiche Monographien zu Theorie und Anwendungen von Punktprozessen. Eine kleine Auswahl findet sich in der Liste am Ende der Arbeit, siehe insbesondere Brémaud [2], Daley and Vere-Jones [5], Jacobsen [9], Karr [11] sowie Last und Brandt [14]. Für Anfänger sind sie mitunter nicht einfach zu verstehen, da sie vom Leser ein hohes technisches Verständnis von stochastischen Prozessen voraussetzen. Das Buch von Snyder [18] bildet eine Ausnahme, da es versucht, an Hand von Anwendungsbeispielen den Leser für die Materie zu gewinnen. Monographien wie Andersen et al. [1] oder Fleming und Harrington [6] sind auf Anwendungen in der Survival Analysis spezialisiert. Shorack und Wellner [17] ist nach wie vor ein Klassiker über empirische Verteilungsfunktionen. Ein Buch, welches neben Rechtszensierungen auch andere Zensierungsmechanismen ausführlich diskutiert, ist mir nicht bekannt. Gleiches gilt für statistische Verfahren bei sich selbsterzeugenden Phänomenen.

8 Schlusskommentar

Ziel dieses Artikels war es, anhand von Beispielen den Bogen zu spannen von sehr einfachen Punktprozessen bis hin zu sich selbsterzeugenden Phänomenen und dabei die wichtigsten Bausteine der Modellierung kennenzulernen. Da die von der

Wahrscheinlichkeitstheorie geprägten Begriffe in Anwendungen tatsächlich unbekannt sind, ist es Aufgabe der Statistik, die in Daten verborgene Information sichtbar zu machen. In vielen Situationen sind Daten verzerrt, zensiert oder trunciert, was neue speziell darauf zugeschnittene Techniken erfordert. Eine seriöse individualisierte Modellierung von Punktprozessen erfordert in der Regel eine Kooperation mit einschlägigen Fachleuten. Insgesamt ein Feld, auf dem zu arbeiten sehr spannend sein kann.

Literatur

1. Andersen, P.K., Borgan, O., Gill, R.D., Keiding, N.: Statistical Models Based on Counting Processes. Springer, New York (1992)
2. Brémaud, P.: Point Processes and Queues: Martingale Dynamics. Springer, Berlin (1981)
3. Cox, D.R.: Regression models and life-tables (with discussion). J. R. Stat. Soc. B **34**, 187–220 (1972)
4. Dabrowska, D.M.: Kaplan-Meier estimate on the plane. Ann. Stat. **16**, 1475–1489 (1988)
5. Daley, D.J., Vere-Jones, D.: An Introduction to the Theory of Point Processes. Springer, New York (1988)
6. Fleming, T.R., Harrington, D.P.: Counting Processes and Survival Analysis. Wiley, New York (1991)
7. Gill, R.D., van der Laan, M.J., Wellner, J.A.: Inefficient estimators of the bivariate survival function for three models. Ann. Inst. Henri Poincaré B, Probab. Stat. **31**, 545–597 (1995)
8. Hawkes, A.G.: Spectra of some self-exciting and mutually exciting point processes. Biometrika **58**, 83–90 (1971)
9. Jacobsen, M.: Statistical Analysis of Counting Processes. Lecture Notes Statist., Bd. 12. Springer, New York (1982)
10. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. **53**, 457–481 (1958)
11. Karr, A.F.: Point Processes and Their Statistical Inference. Dekker, New York (1991)
12. Kopperschmidt, K., Stute, W.: Purchase timing models in marketing: a review. AStA Adv. Stat. Anal. **93**, 123–149 (2009)
13. Kopperschmidt, K., Stute, W.: The statistical analysis of self-exciting point processes. Stat. Sin. **23**, 1273–1298 (2013)
14. Last, G., Brandt, A.: Marked Point Processes on the Real Line: The Dynamic Approach. Springer, New York (1995)
15. Ryan, T.P., Woodall, W.H.: The most-cited statistical papers. J. Appl. Stat. **32**, 461–474 (2005)
16. Sen, A., Stute, W.: The Multivariate Kaplan-Meier Estimator (2013, Zur Publ. eingereicht)
17. Shorack, G.R., Wellner, J.A.: Empirical Processes with Applications to Statistics. Wiley, New York (1986)
18. Snyder, D.L.: Random Point Processes. Wiley, New York (1975)
19. Stute, W.: Kaplan-Meier integrals. In: Handbook of Statistics 23, 87–104 (2004)

Winfried Stute geb. 1946, ist Professor für Mathematische Stochastik an der Justus-Liebig-Universität Gießen. Er studierte an der Ruhr-Universität Bochum und wurde dort 1975 promoviert. Er habilitierte sich 1980 an der LMU München. Von 1981–1983 war er als Professor am Mathematischen Institut der Universität Siegen tätig, seit 1983 in Gießen. Im Jahre 1990 wurde er zum Fellow des Institute of Math. Statistics ernannt, 2008 erhielt er von der Universität Santiago de Compostela die Ehrendoktorwürde verliehen. Seine Hauptarbeitsgebiete sind Stochastische Prozesse und ihre Anwendungen in der Survival Analysis, Marktforschung und Finanzmathematik.

Jahresber Dtsch Math-Ver (2014) 115:153–183
DOI 10.1365/s13291-013-0076-0

SURVEY ARTICLE

An Invitation to Lorentzian Geometry

Olaf Müller · Miguel Sánchez

Published online: 19 December 2013

Abstract The intention of this article is to give a flavour of some global problems in General Relativity. We cover a variety of topics, some of them related to the fundamental concept of *Cauchy hypersurfaces*:

(1) structure of globally hyperbolic spacetimes,
(2) the relativistic initial value problem,
(3) constant mean curvature surfaces,
(4) singularity theorems,
(5) cosmic censorship and Penrose inequality,
(6) spinors and holonomy.

Keywords Global Lorentzian geometry · Cauchy hypersurface · Global hyperbolicity · Einstein equation · Initial value problem · CMC hypersurface · Singularity theorems · ADM mass · Cosmic censorship hypotheses · Penrose inequality · Spinors · Lorentzian holonomy

Mathematics Subject Classification Primary 53C50 · 8306 · Secondary 8302 · 83C05 · 83C75

O. Müller (✉)
Fakultät für Mathematik, Universität Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany
e-mail: olaf.mueller@mathematik.uni-regensburg.de
url: http://homepages-nw.uni-regensburg.de/~muo63888

M. Sánchez
Departamento de Geometría y Topología, Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva s/n., 18071 Granada, Spain
e-mail: sanchezm@ugr.es
url: http://gigda.ugr.es/sanchezm/index.html

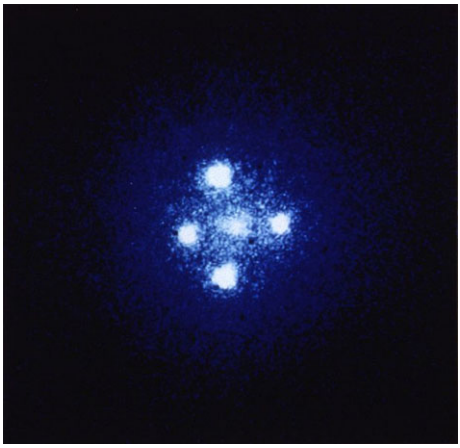


Fig. 1 Einstein's cross, a gravitationally lensed region at 8 billion light years from earth, in the constellation Pegasus. A central galaxy focuses the light emanating from a quasar behind it producing a fourfold image of it. The images can be attributed to a single quasar via comparing characteristic spectra and synchronized relativistic jets. Gravitational lensing is a physical phenomenon closely related to the Lorentzian topic of Morse theory of causal curves [94] (see Sect. 6). Image taken by the European Space Agency's Faint Object Camera on board NASA's Hubble Space Telescope in 1990

1 Introduction

The ground-breaking discovery of the theory of relativity has revealed that the physics of gravity can be described successfully by a theory treating space and time on the same footing, distinguished by the sign of an inner product. Nowadays, the mathematical framework of General Relativity can be regarded as a branch of Geometry (Lorentzian Geometry), in a similar sense like mathematics of Theoretical Mechanics are a branch of symplectic geometry. Admittedly, the physical leading ideas behind the geometric results are more subtle and less evident in General Relativity than in Mechanics. But after getting some familiarity with it, a new geometric world opens up, including unexpected new solutions (and problems) in *Riemannian* Geometry. We want to provide a brief overview of this wonderful world—which might be closer to the reader's field of research than (s)he may expect. We focus on global problems, which are usually the most interesting for mathematicians. We hope that this article will be also of some interest for physicists, which, frequently, are very familiar with local Differential Geometry, while global problems are neglected as of little importance for experimental purposes. Nevertheless, global questions can provide the necessary framework to the full theory and have implications in more practical issues—prominent examples supporting this claim are the definition of mass and energy or the Aharonov-Bohm effect.

One can distinguish different directions in Lorentzian Geometry:

1. *Looking at Riemannian Geometry.* That is, one tries to adapt the familiar Riemannian tools and results to the Lorentzian case, as far as possible. This was the case at the beginning of General Relativity, and is also the typical starting point for a standard mathematician—who has studied Riemannian Geometry but not Lorent-

zian geometry. This is not as straightforward as it sounds, because Lorentzian and Riemannian geometries, in spite of sharing common roots, separate fast in both aims and methods.

2. *Developing specific Lorentzian tools*. Concepts such as causality, boundaries, conformal extensions (Penrose diagrams), asymptotic behaviors (spatial and null infinities) or black holes, are specific to Lorentzian Geometry, without any analog in the Riemannian case. Here, physical intuitions are a very important guide, but we emphasize that these concepts have a completely tidy mathematical definition.
3. *Feed back to Riemannian geometry*. Sometimes, a problem in Lorentzian Geometry admits a full reduction to a purely Riemannian problem. This problem may be unexpected from a Riemannian approach, but now it becomes natural. The initial value constraint equations for Einstein's equation, the positive mass theorems (which yield the last step in the solution to the Yamabe problem!) or the Penrose inequalities provide remarkable examples of this situation. The reader will be able to appreciate that all the main results in Sect. 7 are stated in a purely Riemannian way, even though some motivations in this section, as well as further developments in Sect. 8, show the power and beauty of the bigger Lorentzian world.

In what follows, after a preliminary comparison between Lorentzian and Riemannian Geometries, a short overview of six research areas in Lorentzian Geometry, most of them motivated by General Relativity, is provided. In our choice of problems, global hyperbolicity and Cauchy hypersurfaces play an important role. The reason is twofold: on one hand, they play a central role in global problems, on the other hand, they are a very intuitive bridge between Riemannian and Lorentzian geometry (see Sect. 3). Additionally, the reader will find a variety of further topics, including hyperbolic equations, geodesics, CMC hypersurfaces, mass inequalities, holonomy and spinors. Some parts of this paper extend and update the earlier review [102].

2 From Riemannian to Lorentzian Geometry

In this section, we briefly recall the basics of Lorentzian geometry and compare them to the Riemannian situation (see [17, 28, 48, 83, 89, 114] for further details). Let, throughout this article, M be an n-dimensional manifold, oriented if necessary. A *Lorentzian metric on* M is a symmetric bilinear form of signature $(1, n-1)$ at every point of M, that is, the maximal dimension of a negative definite subspace of T_pM is 1 and the maximal dimension of a positive definite subspace of T_pM is $n-1$. While on any manifold there is a Riemannian metric (using paracompactness), the same is not true any more if one replaces "Riemannian" by "Lorentzian": actually, the existence of a Lorentzian metric on a manifold M is easily seen to be equivalent to the existence of a one-dimensional subbundle of the tangent bundle which implies the vanishing of the Euler class of τ_M, or equivalently, of the Euler characteristic of M [15, Theorem 2.19]. For noncompact manifolds, however, this yields no obstruction: every non-compact manifold carries a Lorentzian metric.

In General Relativity, information is allowed to travel not along arbitrary curves, but only along *causal* ones. The notion of causality is first defined on the tangent

spaces via the sign appearing in the Lorentzian metric. Namely (following the convention in [83]), a non-zero tangent vector $v \in T_pM$ is *causal* if it is either *timelike*, i.e. $g(v,v) < 0$, or *lightlike*, i.e. $g(v,v) = 0$, $v \neq 0$. A vector is called *spacelike* if $g(v,v) > 0$ (in particular, this convention means that $v = 0$ is non-spacelike and also non-causal). Both lightlike vectors and the 0 vector are called *null vectors*.

Let J_g be the set of all the causal vectors and I_g the set of timelike vectors. Then it is easy to see that each of $I_g \cap T_pM$ and $J_g \cap T_pM$ have two connected components. A connected Lorentzian manifold (M, g) is called *time-orientable* if J_g is disconnected; in this case, J_g has exactly two connected components (a fact that can be seen easily by general connectedness arguments using lifts of curves into open fibers). A *time-orientation* is a choice of one component J_g^+, which is called the *causal future* then (and its J_g-complement $J_g^- := -J_g^+$ is called *causal past*). Analogous considerations are valid for I_g. By a *spacetime*, we mean a (connected) time-oriented Lorentzian manifold (M, g) (the choice of time-orientation is not denoted explicitly).

The notion of "future" is then transferred from TM to M by curves: A (continuous) piecewise C^1 curve $c : I \subset \mathbb{R} \to M$ is called *future-directed causal* (resp. *future-directed timelike future*) iff $c'(t) \in J_g^+$ (resp. $c'(t) \in I_g^+$) for all t in each closed subinterval $I_j \subset I$ where c is C^1—and correspondingly for the past. Then, we define the *chronological future* of p as:

$$I_g^+(p) := \{q \in M \mid \exists \text{ future-directed timelike curve } c \text{ from } p \text{ to } q\}.$$

Analogously, the *causal future* $J_g^+(p)$ is defined replacing "timelike" by "causal" in previous definition, and adding by convention $p \in J_g^+(p)$. There are natural dual 'past' notions, and the subscript g is removed when there is no possibility of confusion. Trivially, $I^{\pm}(p)$ are always open, but simple examples show that, in general, $J^{\pm}(p)$ are neither open nor closed—however, $\overline{I^{\pm}(p)} = \overline{J^{\pm}(p)}$ always holds. A state of a classical relativistic system at a point p can then only depend on events in the causal pastof p, and the state at p in turn can only influence the physics in the causal future $J^+(p)$ of p. The distinction of curves by their causal character corresponds to different elements of physical reality: massive bodies (e.g. observers) are supposed to travel along timelike curves, massless particles along lightlike curves, whereas spacelike curves do not admit a direct physical interpretation.

Many constructions in semi-Riemannian geometry are independent of the signature of the metric. First of all, any (non-degenerate) metric determines always a unique metric torsion-free connection (Levi-Civita connection). The Riemannian curvature tensor is defined in exactly the same way, and geodesics are defined as ∇-autoparallel curves; in particular, local convexity (in the sense of existence of a geodesically convex neighborhood around every point) is ensured. In Lorentzian signature, timelike geodesics are local maxima of an appropriately defined length functional (see below) on causal curves, and this property can be extended to lightlike geodesics, even though some relevant subtleties arise (see, for example, [82, Sect. 2]). In the framework of relativistic physics, the "observer" corresponding to a future-directed timelike curve c, is considered to be falling freely when c is a geodesic. Future-directed lightlike geodesics are regarded as "(trajectories of) light rays". No

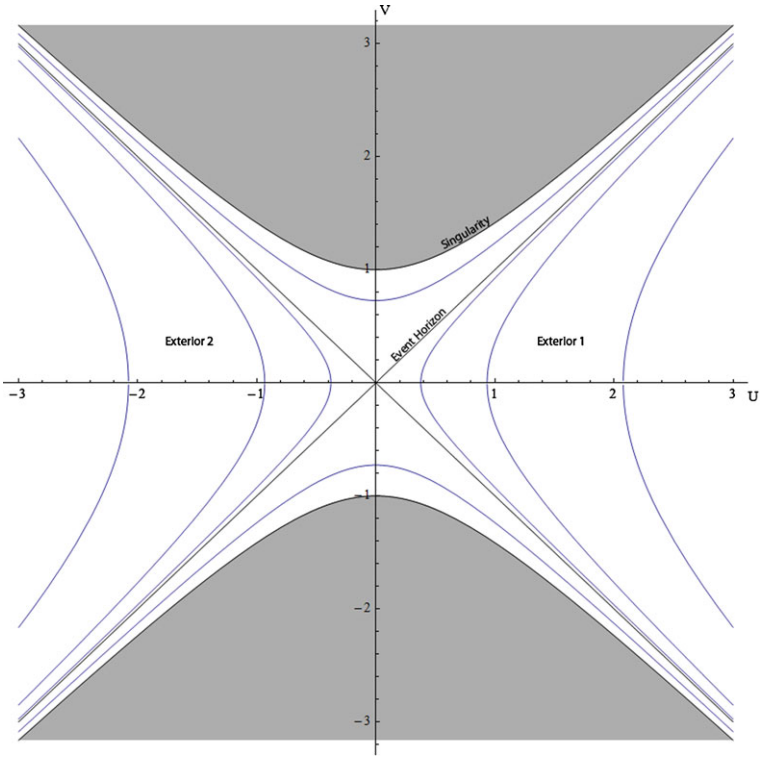


Fig. 2 An instructive example: Schwarzschild-Kruskal metric. The two-dimensional Kruskal metric h is conformally equivalent to the standard Minkowski metric on the complement of the shaded region (in particular, at every point, the future causal cone $J_g^+ \cap T_pM$ is just the upper quadrant between the translates of the diagonals). The conformal factor is $\frac{32M^3}{r}e^{-r/2M}$, and the 4-dimensional spacetime is the warped product $h \times_{r^2} g_{\mathbb{S}^2}$ where r is the unique positive number such that $U^2 - V^2 = (1 - \frac{r}{2M})e^{r/2M}$. Lines through the origin are level sets of the coordinate t which is a time function for the exterior of the event horizon, whereas the displayed hyperbola correspond to constant r coordinate and are Killing orbits, timelike outside and spacelike inside the event horizon. The Schwarzschild part is exactly the future of the line $U = -V$; its exterior part describes successfully the gravitational field of an approximately spherically symmetric mass, e.g. of the sun. All future curves starting in the interior part have finite length (image taken from http://commons.wikimedia.org/wiki/File:KruskalKoords.jpg licensed under the Creative Commons Attribution 3.0 Unported license by Author AllenMcC)

good interpretation holds for spacelike curves, even if they are geodesics, except for very special classes of spacetimes.

Many of the Riemannian constructions carry over to the Lorentzian case *mutatis mutandis*. However, it is worthwhile to give a brief overview of some of the most important differences between Riemannian and Lorentzian geometry.

1. **For any given manifold M, the set of all the Lorentzian metrics on M is not convex**. Recall that the set of Riemannian metrics on M is a (non-empty) convex cone. But this does not hold for Lorentzian metrics, even when there is no topological obstruction for their existence: one can check it trivially in dimension 2 (if g is Lorentzian then $-g$ is Lorentzian too), and that counterexample can easily be

extended to higher dimensions. This fact makes several constructions of interpolation between different metrics significantly more complicated, e.g. in the theory of spinors.

2. **Sectional curvature is defined only for "non-lightlike" planes**. The reason is based on the fact that when the restriction of the metric g to some tangent plane $\pi \subset T_pM$ becomes degenerate, then the denominator in the definition of sectional curvature is 0. This fact also has the consequence that, if the sectional curvatures of the non-degenerate planes at p are not constant, then the sectional curvatures at p will reach values on all $\mathbb{R}$ (recall that one would divide by arbitrarily small positive and negative quantities, depending on the plane), see [89]. Also due to elementary algebraic reasons, a bound of the type $\mathrm{Ric}(v, v) > cg(v, v)$ for all $v \in T_pM \setminus 0$ and some $c \in \mathbb{R}$ cannot hold, and inequalities for curvatures must be understood in the sense of Andersson and Howard, [6]. In particular, the so-called *energy conditions* (see Sect. 4) will play the role of classical curvature bounds in some Lorentzian results.
3. **There is no good embedding of the category of Lorentzian manifolds into metric measure spaces**. Recall that, in order to understand collapse processes and prove finiteness results like the ones due to Cheeger-Gromov, one needs to embed the manifolds and their possible limits into some space of metric spaces, independent of further specializations like Alexandrov spaces, CAT(0) spaces, etc. There is a well-known natural embedding of the category of Riemannian manifolds and convex open isometrical embeddings to the category of metric measure spaces which commutes with the forgetful functor to the category of topological spaces (just by taking the geodesic distance and the associated volume form). That is to say, from a Riemannian metric one can construct in a natural way a metric space compatible with the given topology. In contrast, there is no such construction in the Lorentzian case (which is easily seen by the existence of boosts around p in Lorentz-Minkowski space $\mathbb{L}^n := \mathbb{R}^{1,n-1}$ mapping an arbitrary point in $J^+(p) \setminus I^+(p)$ into an arbitrarily small neighborhood of p). Consequently, all the attempts of repeating the above compactness and finiteness results are doomed to failure in the Lorentzian regime. The same holds for most of the theory of isoperimetic inequalities. See, however, the notion of "Lorentzian distance" below.
4. **The isotropy group of a point may be non-compact**. If a group G acts faithfully and isometrically on a Riemannian manifold (M, g), then the isotropy group of any $p \in M$, being a closed subgroup of $O(n)$, must be compact. This does not hold by any means in the Lorentzian case, as the Lorentz group $O_1(n) \equiv O(1, n-1)$ (which is the isotropy group of any $p \in \mathbb{L}^n$) is not compact. This represents an additional difficulty in finding invariant quantities (which are usually found as an average w.r.t. an associated Haar measure).
5. **No analog to Hopf-Rinow holds**. Geodesic dynamics change drastically: as there is no metric space associated to a given Lorentzian metric, none of the assertions of classical Hopf-Rinow theorem hold in Lorentzian geometry. Geodesic completeness neither implies b.a.-completeness (i.e., the completeness of curves with bounded acceleration) as in the Riemannian case. Neither compactness nor homogeneity of a Lorentzian manifold implies its geodesic completeness (however,

by an argument due to Marsden [80], one knows that both properties together do). A counterexample for homogeneity is just a half-plane $H := \{(x_0, x_1) \in \mathbb{L}^2 | \; x_0 > x_1\}$ (the isometry group include the actions of translations by $\mathbb{R}(\partial_0 + \partial_1)$ plus the so-called boosts, i.e., the connected part of the identity of $O(1,1)$). For compactness, consider the projection of the metric $g := -2dudv + (\cos^4(v) - 1)du^2)$ on the two-dimensional torus $\mathbb{R}^2/\mathbb{Z}^2$. Putting $u := x + t$, $v := x - t$ the g-geodesic $t \mapsto (1/t - t, \arctan(t))$ is then incomplete. Geodesics in Lorentzian manifolds can behave in a strange way for the Riemannian intuition. For example, there are inextensible lightlike geodesics $c : I \to M$ that are closed (in the sense that $c(I) = c(J)$ for a compact subinterval J of I) but they are non-periodic (they appear in the quotient of H above by a discrete subgroup of boosts generated e.g. by $(u, v) \mapsto (2u, v/2)$ where u, v are some natural lightlike coordinates). Moreover, conjugate points along a spacelike geodesic may accumulate [65, 95].

6. **Distinction between irreducibility and indecomposability for isometric actions**, as a Lorentzian vector space is not the sum of a (degenerate) subspace and its orthogonal. Recall that for an isometric Riemannian action ρ on T_pM, the orthogonal of an invariant subspace A of ρ, is invariant itself and complements A. The latter ceases to be true in the Lorentzian case if the metric restricted to A is degenerate. So, the action of a isometry group may be reducible to A, but the vector space T_pM may be not decomposable as sum of irreducible parts. This applies to holonomy representations, and made the discovery of the Lorentzian Berger list (the classification of all possible holonomy groups) considerably more difficult than the one of its Riemannian predecessor. This landmark was completed in 2005, cf. Sect. 8.

At this point, Riemannian geometers should not feel scared off by the above list of differences as they are balanced by a row of nice features listed below. First of all, the Laplacian can be formally defined in the Lorentzian setting as in the Riemannian one, but now it is an hyperbolic operator (d'Alembertian) and, even more, *Lorentzian Geometry is a natural framework to study hyperbolic equations, as Riemannian Geometry is a natural setting for elliptic ones*. This claim is supported by several geometric tools available in the Lorentzian setting but not in the Riemannian one. Let us point out some of these genuinely Lorentzian tools.

1. **Causality allows to visualize the conformal structure of spacetimes**. In fact, the datum $J_g^{\pm}$ which defines the causal structure is conformally invariant and, conversely, two Lorentzian metrics g, g' are (pointwise) conformal (i.e., $g' = \Omega g$, $\Omega^2 > 0$) if they have equal causal cones. One can define the *chronological* $\ll$ and *causal relations* $\leq$, namely $p \ll q$ iff $p \in I^-(q)$, $p \leq q$ iff $p \in J^-(q)$. Locally, either of these relations characterizes the conformal structure. So, *Causality* can be identified to conformal geometry in Lorentzian signature (nevertheless, a subtler modification of the notion of Causality has been recently introduced by García-Parrado and Senovilla [57], see also [56]). Remarkably, one can associate a (conformally invariant) *causal boundary* to every sufficiently well behaved spacetime, this allows to describe the possible asymptotics of timelike curves in a subtle way, ordering them by inclusion of their pasts [50].
2. **Higher compatibility with conformal structures of lightlike pregeodesics**. It is easy to see that there are positive functions of Schwartz class on Euclidean $\mathbb{R}^2$

which do not leave, when used as a conformal factor, any pregeodesic invariant; here, a pregeodesic means a geodesic up to a reparametrization. In contrast, on a Lorentzian manifold, conformal factors leave always some pregeodesics invariant, namely *all the lightlike pregeodesics*. Moreover, conformal changes even preserve conjugate points on them plus their multiplicities, [83].

3. **Completeness and singularities**. Apart from the differences noted above, there are also new notions of completeness in Lorentzian geometry. Traditionally, one distinguishes (logically independent) weaker notions than geodesic completeness: spatial completeness, lightlike completeness and timelike completeness, depending on the causal character of the geodesics in question. Such a subdivision of completeness does not exist in Riemannian geometry and, even more, two stronger notions appear in the Lorentzian setting: the above mentioned *b.a.-completeness* and Schmidt's *b-completeness* [104] (any of them coincide with usual geodesic completeness in Riemannian geometry). As a remarkable difference with the Riemannian case, both, completeness and incompleteness are C^r unstable, for every $r \in \mathbb{N}$, even in the case of Lorentzian metrics on a torus [99], but some results on stability can be still obtained in the Lorentzian case, as C^1-fine stability in the globally hyperbolic case, see [17, 28]. The interplay between these particularities, causality and some physical interpretations, have the effect that, in Lorentzian Geometry, *singularity theorems* (which ensure incompleteness rather than completeness) play an important role, see below.
4. **Reverse triangle inequality and Lorentzian "distance"**. As a construction related to Causality but not conformally invariant, one can define the *time separation* $d(p,q)$ between two points $p, q \in M$ of a spacetime (M, g) as the *supremum* of the lengths of the future-directed causal curves starting at p and ending at q; this supremum may be infinity, and it is regarded as equal to 0 if $p \not\leq q$ (in fact $d(p,q) = 0$ iff $p \not\ll q$). The corresponding function $d : M \times M \in [0, \infty]$ is commonly called the *Lorentzian distance*, as it satisfies a *reverse* triangle inequality (due to the existence of a reverse triangle inequality for causal vectors in the same cone in any T_pM) with some similarity to the Riemannian case. However, it presents also some big differences with the Riemannian case; for example, $d(p,q) > 0$ implies either $d(q,p) = 0$ or there exists a closed timelike curve through p and q (and, so, $d(p,p) = d(q,q) = \infty$). As a clear connection with Causality, $d(p,q) > 0$ iff $p \ll q$ and, under a mild Causality condition on the spacetime (strong causality, i.e., absence of "almost closed" causal curves) the Lorentzian metric g can be reconstructed from d, as in the Riemannian case. However, the interplay of d with Causality makes it a genuinely Lorentzian element. For example, a spacetime is globally hyperbolic (see definition below) iff the Lorentzian distance is finite and continuous for all the metrics in the conformal class of g (see for example [17, 83]).

3 Causality and Global Hyperbolicity

Good conditions on the causality of a spacetime may yield some connections between Riemannian and Lorentzian manifolds and links between hyperbolic and ellip-

tic equations. A key notion is *global hyperbolicity* which is to be developed here and which will play a role in the spirit of *completeness* for Riemannian manifolds.

A spacetime (M, g) is called *globally hyperbolic* iff it is causal[1] (i.e. $p \notin J^+(p)$ for all $p \in M$) and diamond-compact. Here, "diamond-compact" means $J^+(p) \cap J^-(q)$ compact for all $p, q \in M$. The condition of causality corresponds to the existence of global solutions of natural linear differential operators for initial values on maximal achronal hypersurfaces while the condition of diamond-compactness corresponds to their uniqueness: consider the examples of a flat Lorentzian torus for non-existence and of a flat vertical strip in Minkowski space for non-uniqueness. In terms of physics, diamond-compactness corresponds to predictability of nature or Laplace's demon principle (which has been of some influence at least in classical physics), whereas causality corresponds to the exclusion of the possibility of time machines.[2] The link between global hyperbolicity and Riemannian completeness comes from the following result, which lies in the spirit of Hopf-Rinow's (see for example [17, 89]):

Proposition 3.1 *In a globally hyperbolic spacetime* (M, g), *the Lorentzian distance* d *is finite, continuous and satisfies the* Avez-Seifert property, *i.e., for any pair of causally related distinct points* $p, q \in M$ $(p \leq q)$, *there exists a causal geodesic from* p *to* q *with length equal to* $d(p, q)$.

There are many examples of globally hyperbolic manifolds:

1. A Lorentzian product $(\mathbb{I} \times N, -dt^2 + g_N)$ for an interval $\mathbb{I}$, is globally hyperbolic iff g_N is a complete Riemannian metric on N; in particular, Lorentz-Minkowski spaces are globally hyperbolic.
2. *Narrower Cones Principle*: if (M, g) is globally hyperbolic and h is another Lorentzian metric on M with $J_h \subset J_g$, then (M, h) is globally hyperbolic as well; in particular, global hyperbolicity is conformally invariant. This implies that, instead of the Lorentzian products in item 1, one can equally consider *Generalized Robertson-Walker spacetimes* (with complete g_N) i.e, warped products $(\mathbb{I} \times N, -dt^2 + f(t)g_N)$ for some positive function f.
3. If (M, g) is globally hyperbolic and $A \subset M$ is a causally convex open subset of M in the sense that causal curves cannot leave and then re-enter A, then $(A, g|_A)$ is globally hyperbolic as well.
4. Using convex neighborhoods, it is easy to see that any point in any Lorentzian manifold has a globally hyperbolic neighborhood.

[1] Typically, an a priori stronger hypothesis that causality is used to define global hyperbolicity, namely, *strong* causality. In [24] it has been shown that both notions agree. For simplicity, we renounce giving details here and simply use the more recent definition of global hyperbolicity instead.

[2] Time machines are in contradiction to the unspoken fundamental assumption of the free will of the experimentalist taken by the vast majority of physicists, in the sense that any observer, in contrast to physical nature around him, is assumed to be able to take decisions like preparing a spin-up or a spin-down state in a manner which is in principle unpredictable for others, compare the discussion of Bell's inequality and the EPR paradox. Note that without that assumption, time machines do not contradict any other principles of physics—with the possible exception of predictability of nature, cf. the article of Krasnikov [70] as well as its critical reception in [77].

5. Global hyperbolicity of Lorentzian metrics is a C^0-fine stable property in the space of Lorentzian metrics which has been shown by the works of Geroch [64] (taking into account the progress made by Bernal and Sánchez in [22], see [103]); see also Lerner [75] or the extended version on arxiv.org of Benavides and Minguzzi [18].

This last point goes in the direction of Proposition 3.1, i.e. the role of global hyperbolicity is related to Riemannian completeness, as both properties are C^0 stable (but geodesic completeness is not C^r-stable for any r in the general Lorentzian case and only C^1-fine-stable for globally hyperbolic manifolds, as mentioned above).

Geroch [64] showed in 1970 that a spacetime is globally hyperbolic[3] if and only if it contains a Cauchy hypersurface, that is, a set Σ that is crossed exactly once by any inextensible timelike curve (a posteriori, Σ must be then a topological hypersurface, see [89]). Moreover, he gave a construction of a continuous *Cauchy time function* t, which means that t increases strictly monotonously along every causal future-directed curve, and is surjective onto $\mathbb{R}$ along any inextensible causal future-directed curve. His construction involved volumes of sets type $J^{\pm}(p)$ for a finite volume form. For a long time, it was not known, but generally assumed, that Σ could be taken as a smooth and spacelike (non-degenerate) hypersurface and, even more, that for such a prescribed Cauchy hypersurface Σ one could find even a *Cauchy temporal function* vanishing on Σ. This term is more special than the one before and denotes a *smooth* function t whose gradient satisfies $g(\mathrm{grad}t, \mathrm{grad}t) < 0$ with $\mathrm{grad}t$ past-directed, plus the surjectivity property above. (Note that not all smooth Cauchy time functions are temporal: consider, on $\mathbb{L}^2$, the function $t(x_0, x_1) := (x_0 + x_1)^3$.) Functions of this kind automatically lead to metric splitings, that is, they imply that the manifold is isometric to

$$(\mathbb{R} \times N, -f^2 \cdot dt^2 + g_t) \tag{1}$$

where $f > 0$ is a smooth function on $\mathbb{R} \times N$ and g_t is a smooth family of Riemannian metrics on the level sets of t, and all level sets of t are Cauchy. The interest in these questions is obvious: on one hand, (smooth) spacelike Cauchy hypersurfaces are the natural ones for initial data (Einstein equation, Penrose inequality ..., see the next sections); on the other, the orthogonal splitting is useful for many properties: Morse Theory, quantization, to find global coordinates, etc. Moreover, it also leads to remarkable analytic results: not only adapted linear symmetric hyperbolic systems (that is, those given by a first-order differential operator on a vector bundle whose symbol is positive-definite exactly on I_g), enjoy global existence and uniqueness for arbitrary smooth initial values at a Cauchy hypersurface, but the same is true for appropriate nonlinear equations like Yang-Mills equations, as shown by Chrusciel and Shatah [40]. Physically, each temporal function t determines in a natural way not only a one-parameter family of diffeomorphic "physical spaces" (the slices $t =$ constant), but also a Wick rotation, obtained by inverting the sign on $\mathbb{R} \cdot \mathrm{grad}t$ and leaving the orthogonal complement unchanged.

[3] He considered a different, but equivalent, notion of global hyperbolicity, based on the compactness of the space of causal curves connecting each two points, but this is not specially relevant at this point.

Sachs and Wu [100, p. 1155] posed the existence of a *smooth* Cauchy hypersurface in any globally hyperbolic spacetime as a first open "folk" problem. Such a type of problems cannot be overlooked by physicists as minor questions of mathematical rigor, as the requirements in the definition of global hyperbolicity are plausible from the physical viewpoint, but the assumption of a splitting *a priori* of the spacetime as in (1) (the type of expression truly useful for several physical purposes) would be totally unjustified. In a series of papers published along 2003–2006, Bernal and Sánchez [21–23] (see also [101]) gave a full solution by showing that a splitting as (1) can be obtained and, then, any prescribed spacelike Cauchy hypersurface can be chosen as the level $t = 0$ of the splitting; their proof used local convex coordinates patched together in a sophisticated way. There has been, however, quite a few of interesting developments since then. In 2011, Müller and Sánchez [88] solved the question of which Lorentzian manifolds are isometrically embeddable in some $\mathbb{L}^n$ (in the spirit of Nash's theorem). With this aim, they proved, in particular, that any globally hyperbolic spacetime admits a splitting as in (1) with an upper bounded function $f < 1$ (this yielded directly the isometric embeddability of all globally hyperbolic spacetimes). Further properties on both, the splitting (bounds for curvature elements of the slices, flexibility) and the isometric embedding in $\mathbb{L}^N$ (closedness) were obtained then by Müller [86, 87]. In 2012, Fathi and Siconolfi [49] proved the existence of a Cauchy temporal function in a class of geometric spaces with a *cone structure* (which generalized notably the class of globally hyperbolic spacetimes); their proof involves tools from weak KAM theory. By taking into account the progress along these decades (including old work by Seifert), Chrusciel, Grant and Minguzzi [38] have proved very recently that, for some appropriate non-canonical choice of a volume form, also the original functions defined by Geroch become C^1 (and can be smoothed out further by local convolutions). The interplay among these tools is an exciting matter of study [103].

Summing up from a broad perspective, classical elementary results as Proposition 3.1, deeper structural results as splitting (1), and links with other parts of Differential Geometry or Mathematical Physics (Morse theory, Geometric Analysis, Cosmic Censorship, Wick rotation, Einstein and Yang Mills equations ...), show that Riemannian geometry is an indispensable tool in the theory of globally hyperbolic manifolds, but the study of the interplay between the two regimes has just been initiated.

4 Initial Value Problem

Einstein's field equation can be written (in suitable units) as

$$\mathrm{Ric} - \frac{1}{2} S g = 8\pi T. \tag{2}$$

Here, the geometric terms on the left hand side (Ricci tensor Ric, scalar curvature S) are related to a symmetric 2-tensor on the right hand side, the "stress-energy" T, which describes the distribution of matter/energy.

More properly, we must emphasize that the unknown quantity is not only the metric g (with Ric and S): equations for T must be added to get a coupled system with (2). Nevertheless, we will assume for simplicity (in addition to $\dim(M) = 4$, when necessary) the following cases:

- Along this section, $T = 0$ (vacuum), i.e. (2) becomes $\mathrm{Ric} \equiv 0$.
- In the next sections, solutions with T non-determined but satisfying only any of the (mild) "energy conditions" as: (1) Weak: $T(v, v) \geq 0$ for any timelike v (density energy is nonnegative), (2) Dominant: $-T(v, \cdot)^\flat \equiv -g^{ij} T_{jk} v^k$ is either future-directed causal or 0 for any future timelike v (energy flow is causal), (3) Strong: equivalent via Einstein equation to the timelike convergence condition, $\mathrm{Ric}(v, v) \geq 0$ for timelike v (gravity, on average, attracts).

The well-posedness of Einstein equation requires an input of initial data on a 3-manifold Σ which permits to obtain a ("unique, maximal") spacetime (and eventually a T) such that Σ is embedded in M consistently with the initial data. The problem is complicated: a classical theorem such as Cauchy-Kovalevskaya's is not applicable, and, even more, in principle the system of equations is not hyperbolic. Nevertheless, there exist a highly non-trivial procedure—based on the existence of *harmonic coordinates*—which allows one to find an equivalent (quasi-linear, diagonal, second order) hyperbolic system. The standard global result was obtained by Choquet-Bruhat and Geroch [35]:

Theorem 4.1 *Let* (Σ, h) *be a* (*connected*) *Riemannian* 3*-manifold, and* σ *a symmetric two covariant tensor on* Σ *which satisfies the compatibility conditions of a second fundamental form* (*Gauss and Codazzi eqns.*) *Then there exist a unique spacetime* (M, g) *satisfying the following conditions*:

(i) $\Sigma \hookrightarrow M$, *consistently with* h, σ (*i.e.*, $h = g|_\Sigma$ *etc.*)
(ii) *Vacuum*: $\mathrm{Ric} \equiv 0$ (*this can be extended to any family* T *of natural divergence-free symmetric* 2*-tensors, e.g. to those coming from natural symmetric hyperbolic field theories*).
(iii) Σ *is a Cauchy hypersurface of* (M, g).
(iv) *Maximality*: *if* (M', g') *satisfies* (i)–(iii) *then it is isometric to an open subset of* (M, g).

As suggested previously, the property (iii) becomes essential for the well-posedness of the problem—namely, the existence of a solution spacetime can be proven because no timelike curve crosses Σ twice, and the uniqueness because all timelike curves cross Σ at least once.

Remark 4.2 (SCCC) Even though the solution (M, g) provided by Theorem 4.1 is maximal, it may be extensible as a spacetime, that is, (M, g) may be isometric to an open proper subset of another spacetime $(\bar{M}, \bar{g})$—even a vacuum one. In this case, Σ cannot be a Cauchy hypersurface of the extension, and two possibilities arise: (a) $(\bar{M}, \bar{g})$ is not globally hyperbolic or (b) the initial Σ was not "chosen adequately", as an input hypersurface for a whole physically meaningful spacetime. Thus, an important question is how to characterize the (in)extendibility of (M, g).

This question becomes extremely important in General Relativity because physical intuition suggests that spacetime is inextensible, but it suggests at the same time that it should be predictable from initial data and, thus, globally hyperbolic.

The *Strong Cosmic Censorship Conjecture (SCCC)* asserts that, for generic physically reasonable data (including a "good choice" of Σ), (M, g) is inextendible. Of course, a non-trivial problem of the conjecture, is to explain carefully what "generic physically reasonable data" means.

A systematically studied problem is to characterize/classify the solutions of (vacuum) Einstein equation. By using Theorem 4.1, this is rather a purely Riemannian problem (roughly: given data as, say, (Σ, h), classify the σ's which satisfy Gauss and Codazzi equations). There are two specially important methods of solution (see [11] for a detailed exposition and [37] for updated references):

- *Conformal*. Initial data are divided into two subsets: a subset of *freely specified* conformal data (the conformal class of h, a scalar field τ, and a symmetric divergence free 2-tensor $\tilde{\sigma}$), and a subset of *determined* data (a function $\phi > 0$, a vector field $W \in \chi(\Sigma)$), which are derived from the free data by means of differential equations. The interpretation and equations for these data vary with two types of conformal method (the method (A) or semi-decoupling, whose origin goes back to Lichnerowicz [76], and the method (B) or conformally covariant). The problem is then to show if there exists solutions for the equations of the determined data, and classify them.
- *Gluing solutions*. As a difference with the conformal method, this is not a general one, but it is very fruitful in relevant particular cases. Corvino and Schoen [41, 42] glue any bounded region of an asymptotically flat spacetime with the exterior region of a slice of Kerr's—this case becomes specially interesting as the "no hair theorems" highlight Kerr spacetime at the final state of the evolution of a black hole. The useful gluing by Isenberg et al. ([68], see also the initial data engineering [39]) constructs consistent initial data for Einstein equation from the connected sum of previously obtained data (for example, construction of wormholes).

For the general conformal method, the results depend on different criteria—topology of Σ, asymptotic behaviour, regularity (analytic, smooth, Hölder class ...), metric conformal class (Yamabe type) ... . The most important one is the mean curvature H. Essentially, when H is constant almost all is known (at least if $T = 0$); in fact, if Σ is either compact without boundary, or asymptotically flat or hyperbolic, it is completely determined which solutions exist (and they exist for all but certain special cases). When H is nearly constant there are many results, but also many open questions; otherwise, there are very few results.

5 Constant Mean Curvature Spacelike Hypersurfaces

The importance of (spacelike) hypersurfaces of constant mean curvature (CMC) H in a spacetime (M, g), has been stressed above in relation to the initial value problem, but they are also important for other issues in General Relativity (see the survey by

Marsden and Tipler [79]). Here we will explain some results on their existence and uniqueness. We will restrict to the case when the hypersurface Σ is spacelike, as it is easy to see that a submanifold can extremize area only when its dimension is equal either to the index or to the coindex of the metric (otherwise, the area may be critical, but not extremal). Moreover, when $H = 0$ the spacelike hypersurfaces are either maximal or neither maximize or minimize area. Nevertheless, in both cases they are usually called *maximal*, just like Riemannian minimal surfaces. About the results on existence, we point out (see Gerhardt's book [62] for a detailed study):

- After the special case of Lorentz Minkowski (see below), the first natural problem to be considered is the construction of one CMC hypersurface or, in general, a spacelike hypersurface Σ with a *prescribed* mean curvature H, in a given spatially compact globally hyperbolic spacetime. A relevant result of global existence was due to Claus Gerhardt [59] in 1983, under the condition of existence of barriers. An upper resp. lower r-barrier is a closed spacelike achronal hypersurface with mean curvature $> r$ resp. $< r$. If there is an upper r-barrier Σ^+ and a lower r-barrier Σ^-, Gerhardt shows that there is a CMC hypersurface of mean curvature r in $I^+(\Sigma^-) \cap I^-(\Sigma^+)$. This is shown by solving the Dirichlet problem for a given boundary curve by making some a priori gradient estimates and, then, by applying a Leray-Schauder fixed point theorem. To enhance the constructiveness in the last part, Ecker and Huisken [45] used an evolutionary equation in terms of the mean curvature flow starting at some Cauchy hypersurface $\Sigma \subset I^+(\Sigma^-) \cap I^-(\Sigma^+)$ which provided a better control over the hypersurfaces—for example, it allows to fix all points of vanishing mean curvature during the process. Even though they had to assume some additional conditions (the timelike convergence condition and a more technical structural monotonicity condition), Gerhardt [60] refined Ecker and Huisken's flow method, showing that such additional conditions were unnecessary. The improved control allows to solve also related problems such as: (a) given a prescribed point in M, construct a CMC hypersurface passing through the point, or (b) given a compact Cauchy surface in M find a compact CMC Cauchy surface with the same volume.
- Removing spatial compactness, the next step is to consider the existence of CMC hypersurfaces in asymptotically flat spacetimes. Substantial contributions, specially in the maximal (eventually up to a compact subset) case, have been made by Bartnik [9], Bartnik, Chrusciel and Murchadha [10], and Ecker [44]. All three articles assume an energy inequality and, moreover, a connection between radial and time variables (called "uniform interior condition" in the first and "bounded interior geometry" in the second article). Roughly, both versions of the latter condition assert that the deviations from Minkowski geometry propagate with subluminal velocity—so one can expect the condition to be true for massive Klein-Gordon Theory, e.g. Again, the first article uses Leray-Schauder's fix point theorem, while the other articles use long-time convergence of the mean curvature flow, which provides a better control on the surfaces.
- The question of constructing a whole foliation by CMC hypersurfaces was also studied. In the cited 1983 article by Gerhardt [59], he considered a globally hyperbolic spacetime with compact Cauchy hypersurfaces satisfying the timelike convergence condition. Under these hypotheses, slices with a CMC $H \neq 0$ are unique

and, if there are two different maximal slices, then both have to be totally geodesic and the region enclosed by them must be static—thus, the existence of two different hypersurfaces of CMC implies strong obstructions. In fact, the mean curvature must increase monotonously in foliations by CMC hypersurfaces. The timelike convergence condition is replaced by the mere assumption of a lower bound to the Ricci tensor on timelike vectors in a second article by Gerhardt [61]. This article also treats exclusively the case of globally hyperbolic spatially compact spacetimes. Here, the statement is the following: If for some sequence of Cauchy surfaces Σ_n of M with $\Sigma_{n+1} \subset I^+(\Sigma_n)$ for all n and $\bigcap I^-(\Sigma_n) = M$ there is a sequence of n-barriers $B_n \subset I^+(\Sigma_n)$ for any $n \in \mathbb{N}$, then there is a Cauchy surface Σ of M such that $F := I^+(\Sigma)$ can be foliated by a CMC foliation and the mean curvature is a temporal function on F.

In the last item, some results on uniqueness of CMC hypersurfaces appear implicitly, but this question deserves a bigger attention. A neat problem on uniqueness can be stated as follows, see [1]. Consider a Riemannian n-manifold (M, g_R), a smooth positive function defined on some interval $f : I \subset \mathbb{R} \to \mathbb{R}$, and the solutions u to the differential equation on M:

$$u(M) \subset I, \qquad |\nabla u| < f(u)$$
$$\operatorname{div}\left(\frac{\nabla u}{f(u)\sqrt{f(u)^2 - |\nabla u|^2}}\right) = nH - \frac{f'(u)}{\sqrt{f(u)^2 - |\nabla u|^2}}\left(n + \frac{|\nabla u|^2}{f(u)^2}\right) \tag{3}$$

for some constant H. The graphs of its solutions can be regarded as the spacelike[4] hypersurfaces of CMC equal to H in a Generalized Robertson-Walker spacetime $I \times_f M$ (recall, abusing of the notation $g \equiv -dt^2 + f(t)^2 g_R$). Notice that this *Calabi-Bernstein equation* is the Euler Lagrange one for the functional

$$\mathcal{A}(u) = \int_M f(u)^{n-1}\sqrt{f(u)^2 - |\nabla u|^2}\, dV$$

under the constraint $\int_M (\int_{u_0}^u f(t)^n) dV =$ constant. A specially relevant case of this equation was solved by Cheng and Yau [34] (after the solution by Calabi for $n \leq 4$):

Theorem 5.1 *The only entire solutions to Calabi-Bernstein equation* (3) *in* $\mathbb{L}^{n+1}$ (*i.e.*, $(M, g_R) \equiv \mathbb{R}^n$, $I \equiv \mathbb{R}$, $f \equiv 1$) *are linear* (*or affine*) *functions.*

As a consequence, the only complete maximal hypersurfaces in Lorentz-Minkowski space are the spacelike hyperplanes.

In fact, they proved that any maximal spacelike hypersurface which is also a closed subset in $\mathbb{L}^{n+1}$ is a hyperplane. This yields a surprisingly simple solution to the Calabi-Bernstein problem (recall, for example, that the analogous results change dramatically for minimal hypersurfaces with dimension larger than seven).

[4] Because of the gradient condition $|\nabla u| < f(u)$.

There were, however, well-known counterexamples for the case of CMC hypersurfaces with $H \neq 0$, [110, 112]. So, a line of results about the uniqueness of CMC hypersurfaces has appeared. By using integral inequalities, one can check that all compact CMC hypersurfaces Σ in any GRW spacetime under the *null convergence condition* (i.e., $\mathrm{Ric}(v, v) = 0$ on null vectors) are totally umbilical—and thus, under mild conditions, Σ is a slice $t =$ constant. Such a type of result can be also generalized further, see [1, 84] and references therein. The feed-back of these results with Riemannian ones have been especially fruitful. Remarkably, both the GRW structure and the restriction for the hypersurfaces of being *spacelike*, yields simplifications that inspired some hypotheses for the Riemannian case.

Under some conditions, previous results can be extended to the case when Σ is complete but non-compact, [4]. In general, for the non-compact case, the so-called *Omori-Yau maximum principle* (or *asymptotic Cheng-Yau principle*), becomes useful. This principle is stated for complete, connected, noncompact Riemannian manifolds and, roughly speaking, means that any smooth function u bounded from above on M, will admit a sequence $\{x_k\} \subset M$ which plays the role of a maximum (say, $\lim_{k\to\infty} u(x_k) = \sup_M u$, $|\nabla u(x_k)| \leq 1/k$ and $\Delta u(x_k) \leq 1/k$). The principle holds when the Ricci curvature is bounded from below as well as in other more refined cases. We refer to [3, 96] and references therein for the recent progress on the Omori-Yau principle and its applications to hypersurfaces in both, Riemannian and Lorentzian Geometry.

It is also worth pointing out that, when $\dim(M) = 3$ (i.e., the hypersurface Σ is a surface), new tools appear. For example, a different approach to the non-compact case has been developed very recently for CMC spacelike surfaces in certain 3-dimensional GRW spacetimes $I \times_f F$; the main idea is to prove that, under some natural assumptions, a metric conformal to the induced one on the surface Σ must be parabolic, see [98] and references therein. Recall also that the analog to the classical Björling problem (construct a minimal surface in $\mathbb{R}^3$ containing a prescribed analytic strip, solved by H.A. Schwarz in 1890) has been also considered in the Lorentzian case; this yields a representation formula for maximal surfaces and allows to construct new ones explicitly; see [2, 33] for the case of $\mathbb{L}^3$ and [85] for the general problem in arbitrary spacetimes, without restriction of the dimension.

6 Geodesics and Singularity Theorems

In some concrete spacetimes, singularities might be defined "by hand" but a general definition is difficult [63], for example:

1. The singularity will not be a point of the spacetime, but placed "at infinity"—but no natural notion of infinity exists in general.
2. The curvature tensor R is expected to diverge, but all its scalar invariants ($\sum R_{ijkl}R^{ijkl}$, $\sum \nabla_s R_{ijkl}\nabla^s R^{ijkl}$, $S \ldots$) may vanish when $R \neq 0$.

At any case, some sort of "strange disappearance" happens if the spacetime is *inextensible, but an incomplete causal geodesic exists*, and these two conditions will be regarded as *sufficient* for the existence of a singularity. Then, the aim of the so-called

singularity theorems is to prove that causal incompleteness occurs under general natural conditions on T (an energy condition) and on the causality of the manifold, as global hyperbolicity. Nevertheless, recall that, rather than "singularity" results, they may be "incompleteness" ones: the physical conclusion of these theorems could be that a physically realistic spacetime cannot be globally hyperbolic, rather than being singular. So, they become "true singularity" results when an assumption as global hyperbolicity is removed ... or if SCCC (Remark 4.2) is true!

Recall the following Hawking's singularity theorem (see [48] or [89] for a detailed exposition):

Theorem 6.1 *Let* (M, g) *be a spacetime such that*:

1. *It is globally hyperbolic*.
2. *Some spacelike Cauchy hypersurface* Σ *strictly expanding* $H \geq C > 0$ (H *is the future mean curvature and expansion means "on average"*).
3. *Strong energy* (*i.e.*, *timelike convergence condition*) *holds*: $Ric(v, v) \geq 0$ *for timelike* v.

Then, *any past-directed timelike geodesic* γ *is incomplete*.

Sketch of proof The last two hypotheses imply that any past-directed geodesic ρ normal to Σ contains a focal point if it has length $L' \geq \frac{1}{C}$. Thus, once Σ is crossed, no γ can have a point p at length $L > \frac{1}{C}$ (otherwise, a length-maximizing timelike geodesic from p to Σ with length $L' \geq L$ would exist by global hyperbolicity, a contradiction). □

This result is very appealing from a physical viewpoint, because the assumption on expansion seems completely justified by astronomical observations. Remarkably, the hypothesis $H \geq C > 0$ for some constant C cannot be weakened into $H > 0$, as shown by a surprising (as physically realistic and far-from-vacuum) example due to Senovilla [108]. From a mathematical viewpoint, the reader can appreciate the isomorphic role of the hypotheses above with the typical ones in Myers type results, say: global hyperbolicity/ (Riemannian) completenes and timelike convergence condition/positive lower bound on the Ricci tensor.

Singularity theorems combine previous ideas with (highly non-trivial) elements of Causality. Essentially, there are two types:

1. Proving the existence of an incomplete timelike geodesics in a global, cosmological setting.

   This is the case of Theorem 6.1, and some hypotheses there (specially glob. hyp.) are weakened or replaced by others. For example, Hawking himself proved that, if Σ is compact, global hyperbolicity can be replaced by assuming that Σ is achronal (i.e., non-crossed twice by a timelike curve). In this case, the timelike incompleteness conclusion holds, but in a less strong sense: at least one timelike incomplete geodesic exist.
2. Proving the existence of an incomplete lightlike geodesic in the (semilocal) context of gravitational collapse and black holes.

For the latter, the notion of (closed, future) *trapped* surface K (or $n-2$ submanifold) becomes fundamental. Its mathematically simplest definition says that K is a compact embedded spacelike surface without boundary, such that its mean curvature vector field $\vec{H}$ is future-directed and timelike on all K [107]—essentially, this means that the area of any portion of K is initially decreasing along *any* future evolution; when it is only non-increasing, K will be said *weakly trapped*. Trapped surfaces are implied by spherical gravitational collapse. One would expect that, at least in asymptotically flat spacetimes (see next section), they must appear if enough matter is condensed in a small region and, under suitable conditions, must imply the existence of a *black hole* (see [43] and references therein). That is, the physical claim is that "gravitational collapse implies incompleteness", and a support for this claim is provided by the following Penrose's theorem (the first modern singularity theorem [91]—after the works by Raychaudhuri and Komar):

Theorem 6.2 *Let* (M, g) *be a spacetime such that*:

1. *Admits a non-compact Cauchy hypersurface.*
2. *Contains a trapped surface.*
3. $Ric(k,k) \geq 0$ *for lightlike* k.

Then there exist an incomplete future-directed lightlike geodesic.

As emphasized by Senovilla [109], the pattern of a singularity theorem has three ingredients: firstly, a bound on the Ricci curvature, secondly, a causality condition, and thirdly, an initial condition on a nonzero-codimensional subset. Remarkably, a unified treatment of both types of singularity theorems has been carried out recently by Galloway and Senovilla [55]. Singularity theorems are very accurate, even though it would be desirable to obtain general results on the nature of the incompleteness, or ensuring divergences of R in some natural sense. So, the finding of further types of singularity theorems would be very desirable for physical purposes [108, 109].

Remark 6.3 The subtleties of Lorentzian completeness also appear in the Lorentzian analogue of Cheeger-Gromoll theorem (see for example [17, Chap. 14]). To obtain the Lorentzian splitting, of a spacetime (M, g), $\dim(M) > 2$, as a product $(\mathbb{R} \times \Sigma, -dt^2 + g_\Sigma)$, where (Σ, g_Σ) is a complete Riemannian manifold, one imposes: (a) *either geodesic completeness or global hyperbolicity*, (b) the *timelike convergence condition* (as the meaningful weakening of the positive semi-definite character of the Ricci tensor in the Riemannian case), and (c) the existence of a complete *timelike* geodesic line.

It is also worth pointing out that the variational approach for Riemannian geodesics can be extended to the Lorentzian setting but important particularities appear. It is known since the old work by Uhlenbeck [113] that Morse theory can be applied to lightlike geodesics, under some conditions (including in particular global hyperbolicity). Lightlike geodesics satisfy also a relativistic *Fermat principle* [69, 94]. Combining both facts, one can study *gravitational lensing*, that is, the reception at some point p of the spacetime of light rays arriving in different directions from

the same stellar object, the latter represented by some timelike curve c, see [92, 93]. As shown in [29], a very precise result on the existence and multiplicity of light rays from c to p in physically realistic spacetimes, can be stated in terms of the geodesics connecting two points for an appropriate Finsler metric. For the variational study of spacelike geodesics, see [28, 81] and references therein.

7 Mass, Penrose Inequality and CCC

Asymptotically flat 4-spacetimes are useful to model the spacetime around an isolated body. They can be defined in terms of Penrose conformal embeddings, even though the definition is somewhat involved (see for example [52, 114]). Nevertheless, in what follows it is enough to bear in mind that, in an asymptotically flat (4-)spacetime there exists spacelike Cauchy hypersurfaces Σ which admits an *asymptotically flat* chart $(\Sigma\backslash K, (x_1, x_2, x_3))$ as follows. For some compact $K \subset \Sigma$ and some closed ball $\overline{B_0(R)}$ of $\mathbb{R}^3$, $\Sigma\backslash K$ is isometric to $\mathbb{R}^3\backslash\overline{B_0(R)}$ endowed with the metric:

$$h_{ij} - \delta_{ij} \in O(1/r), \qquad \partial_k h_{ij} \in O(1/r^2), \qquad \partial_k \partial_l h_{ij} \in O(1/r^3), \tag{4}$$

in Cartesian coordinates (this means that Σ is intrinsically asymptotically flat, as a Riemannian 3-manifold; in particular, Ric and S, are in $O(1/r^3)$), and, even more, its second fundamental form σ satisfies: $\sigma_{ij} \in O(1/r^2)$, $\partial_k \sigma_{ij} \in O(1/r^3)$. (This definition can be extended to include more than one end, each one isometric to $(\Sigma\backslash K, (x_1, x_2, x_3))$ as above.)

The total ADM (Arnowit, Deser, Misner) *mass of an asymptotically flat Riemannian 3-manifold* can be defined as the limit in any asymptotic chart:

$$m = \frac{1}{16\pi} \lim_{r\to\infty} \sum_{i,j=1}^{3} \int_{S_r} \left(\partial_i h_{ij} - \partial_j h_{ii}\right) n^j dA, \tag{5}$$

where n is the outward unit vector to S_r, the sphere of radius r. Notice that m depends only on the Riemannian 3-manifold; in fact, when this manifold is seen as a hypersurface of an asymptotically flat spacetime, the appropriate name for m is *ADM energy*, and the definition of mass depends on σ, see the next section. This definition of mass is not mathematically elegant, but recall:

1. ADM mass appears naturally in a Hamiltonian approach, as an asymptotic boundary term for the variations of $\int S$. The definition is not trivial because no strictly local notion of relativistic energy is available—nevertheless, it is worth pointing out the attempts to define a quasilocal mass [111].
2. There exists a classical Newtonian analog when the spacetime is Ricci-flat outside $\mathbb{R} \times K$, K compact, and there exists a timelike Killing vector field ξ with $\lim_{r\to\infty} |\xi| = 1$, such that $\Sigma \perp \xi$. In this case, the divergence theorem yields:
   $m = \frac{1}{4\pi} \int_K |\xi|^{-1} \mathrm{Ric}(\xi, \xi) dV = \int_K |\xi| \rho dV$
   i.e., the "integral of the poissonian density ρ measured at ∞".

3. The expression in coordinates for m is manageable:
   - If $h_{ij} = u^4\delta_{ij}$ with $u(x) = a + \frac{b}{|x|} + O(\frac{1}{|x|^2})$ then $m = 2ab$.
     In particular, this is the case if u is "harmonically flat" i.e. harmonic with finite limit at ∞.
   - Otherwise, when $S \geq 0$ then h is perturbable to the harmonically flat case with arbitrarily small error for m and preserving $S \geq 0$ (Schoen and Yau [106]; Corvino [41] extended the result for $m > 0$ without error in the mass).
4. Classical outer Schwarzschild metric can be written as:
   $M = \mathbb{R} \times \Sigma$, where $\Sigma = \mathbb{R}^3 \backslash \overline{B_0(|m|/2)}$;
   $g = -((1 - \frac{m}{2|x|})/u)^2 dt^2 + h$, $h_{ij} = u^4\delta_{ij}$ with $u = 1 + \frac{m}{2|x|}$
   (in particular $\sigma \equiv 0$). Of course, the classical Schwarzschild mass m agrees ADM mass.

One expects from the physical background that, when the dominant property holds, the ADM mass will be positive for any asymptotically flat Cauchy Σ. Two technical points are relevant here: (a) When Σ is totally geodesic ($\sigma \equiv 0$) the dominant property yields $S \geq 0$. (b) Under our definition of asymptotic flatness, Σ is necessarily complete, but the Riemannian part of exterior Schwarzschild spacetime $(\mathbb{R}^3 \backslash \overline{B_0(|m|/2)}, h)$ is incomplete for any $m \neq 0$. Of course, this is not a problem for the computation of the limit in the expression of the ADM mass, and one can also extend and modify $(\mathbb{R}^3 \backslash \overline{B_0(|m|/2)}, h)$ in a bounded region to obtain a complete Riemannian manifold Σ^c with the same asymptotic behaviour. Moreover, in the globally hyperbolic case $m > 0$, one can obtain such a Σ^c (say, corresponding to the spacetime created by a star of the same mass) with: (i) the same asymptotic behaviour, (ii) $S \geq 0$. Clearly, this property is not expected in the non-globally hyperbolic case $m < 0$. And, in fact, it is forbidden by the *Riemann positive mass theorem*:

Theorem 7.1 *Let (Σ, h) be any asymptotically flat (complete) Riemannian manifold with $S \geq 0$. Then, $m \geq 0$ and equality holds iff (Σ, h) is Euclidean space $\mathbb{E}^3 = (\mathbb{R}^3, \delta)$.*

Remark 7.2 This celebrated result by Schoen and Yau [105] is a purely Riemannian one. From this case, more general "positive mass" results follow, which include the case $\sigma \not\equiv 0$ [106]; see also the comments in the next section about Witten's, completely different, proof. By the way, recall that the solution of Yamabe problem was completed by using the above result (see the nice survey [72]).

It is worth pointing out that, because of a technical problem which goes back to the known failure of regularity of minimal surfaces in dimensions greater than 7, the positive mass theorem is proved for dimensions up to 7 (this has been completed only recently, by using Schoen and Yau techniques, see [47]) as well as for spin manifolds of any dimension (by using Witten's techniques to be explained in the next section).

Next, we will consider a no less spectacular further step (for a detailed exposition, see [27]). But, first two notions will be briefly explained:

1. *WCCC*. A question related with SCCC (see Remark 4.2) is the so-called *weak cosmic censorship conjecture* (WCCC), which is stated in the framework of asymptotically flat spacetimes. In such spacetimes, a natural notion of asymptotic future

null infinity $\mathcal{J}^+$ can be defined ($\mathcal{J}^+$ is a subset of the image of M for a suitable conformal embedding in a bigger spacetime $\bar{M}$) and, then, also a rigorous notion of the *black hole* region B of M appears ($B = M \setminus J^-(\mathcal{J}^+)$)—this region corresponds to the intuitive idea of a "spatially bounded region from where nothing can escape". WCCC asserts that (maybe only generically) any spacetime M obtained as the maximal evolution of physically reasonable initial data with an asymptotic decay,[5] will be asymptotically flat and, in a restrictive sense, globally hyperbolic at infinity.[6] The physical interpretation of this assertion is that no singularity (except at most an "initial" one) can be observed from $M \setminus B$, that is, singularities must lie inside a black hole and cannot be seen from outside (singularities are not "naked").

2. *Outermost trapped surfaces*. Given a totally geodesic asymptotically flat slice Σ, those trapped surfaces (more precisely, compact spacelike surfaces whose expansion respect to the outer future lightlike direction is at no point positive) contained in Σ which are boundaries of a 3-manifold, are known to satisfy:

1. Such trapped surfaces correspond to compact minimal surfaces of Σ.
2. The outermost boundary compact minimal surfaces (necessarily topological spheres, each one the "apparent horizon in Σ of a black hole") are well-defined.
3. Let $\mathcal{H}$ be the union of the outermost minimal surfaces. Under WCCC, if $\mathcal{H}$ is connected and A_0 denotes its area, physical considerations ensure that the "contribution to the mass" m_0 of the corresponding black hole would satisfy: $m_0 \geq \sqrt{\frac{A_0}{16\pi}}$.

Therefore, choosing any asymptotic Σ one expects for its mass m_Σ:

$$m_\Sigma \geq \sqrt{\frac{A_0}{16\pi}} \tag{6}$$

(at least if the second fundamental form vanishes). But (6) is an inequality in pure Riemannian Geometry. Thus, the following precise result must hold:

Theorem 7.3 *Let (Σ, h) be an asymptotically flat Riemannian 3-manifold with $S \geq 0$, and let $\mathcal{H}_0$ be the largest outermost (connected) minimal surface, with area A_0. Then inequality (6) is satisfied, and the equality holds if and only if (Σ, h) is Schwarzschild Riemannian metric outside $\mathcal{H}_0$.*

This is the celebrated "Riemann-Penrose inequality", proved by Huisken and Ilmanen [67] (who re-prove then the Riemann positive mass theorem), and shortly after extended by Bray to the full area of the (maybe non-connected) $\mathcal{H}$, with a different proof [26] based on positive mass theorem.

Penrose inequality is a more general conjecture, which includes the full spacetime case $\sigma \neq 0$ (recall that the case above would correspond when $(\Sigma, h, \sigma = 0)$ can

[5] Typically, this data must satisfy: (i) (Σ, h, σ) is asymptotically flat, (ii) T satisfies the dominant property, and the equations for T constitute a quasilinear, diagonal, second order hyperbolic system, (iii) the fall-off of the initial value of T on Σ is fast enough for the h-distance, and h is assumed to be complete.

[6] More precisely, the latter means that the spacetime is *strongly asymptotically predictable*, see [114]. Recall that WCCC cannot be regarded as a particular case of SCCC.

be regarded as an initial data set for the spacetime). It is still open, and it becomes a major problem in Differential Geometry. An evidence of its difficulty is that it is supported by physical grounds, and counterexamples to more general appealing mathematical conjectures have been found, see [32]; we refer to the reviews [37, 78] for comprehensive references.

8 Spinors and Holonomy

Dirac operators are popular objects of study in the area of global analysis, one of the main reasons being the existence of index theorems for them, see the standard textbook by Lawson and Michelson [71] or the book by Berline, Getzler, Vergne [20] on Dirac operators, both almost exclusively treating the Riemannian situation—this reflects the fact that index theory presently is applicable almost exclusively to elliptic operators. Another main reason of interest in spinors, more predominant in the Lorentzian case, is the presence of *Weitzenböck formulas*. These formulas reflect the fact that the Dirac operator as a natural first-order operator on spinors is a root of the Laplacian type operator plus a curvature-induced zeroth order term.

While in the book by Lawson-Michelson real spinors play a prominent role, in the following, we want to focus here on *complex* spinors.

Spinor bundles are defined verbatim in the same way as in the Riemannian case, with $SO(n)$ always replaced by the connected component of the identity of $SO(1, n)$, but there are some important differences of the Lorentzian to the Riemannian case: The natural (pseudo-Hermitean) scalar product $\langle \cdot, \cdot \rangle$ on the spinor bundle is not definite, but of split signature. Any timelike vector field X can be used to define a (non-natural) positive-definite scalar product $(\cdot, \cdot) := \langle X \cdot, \cdot \rangle$. Clifford multiplication is $\langle\rangle$-symmetric (instead of antisymmetric, as in the Riemannian case).

For a pseudo-Riemannian spin manifold of arbitrary signature, one can define the *Dirac operator* on C^1 (or at least $W^{1,p}$) sections ψ of the spinor bundle by $D\psi := \sum_{i=1}^{n} \epsilon_i e_i \cdot \nabla_{e_i} \psi$ (where the e_i are a pseudoorthogonal basis and ϵ_i is the sign of $g(e_i, e_i)$). The Dirac operator is formally self-adjoint, essentially self-adjoint if (M, g) is complete, and satisfies the Weitzenböck identity

$$D^2 = \nabla^* \nabla + \frac{1}{4} S,$$

where S is the pseudo-Riemannian scalar curvature of (M, g).

In the Riemannian situation, as the connection Laplacian $\nabla^* \nabla$ is positive-definite, the Weitzenböck formula is the initial point of many obstructions to positive scalar curvature for spin manifolds. The Weitzenböck formula for the Lorentzian Dirac operator looks superficially the same but the connection Laplacian here is a *hyperbolic* operator instead of an elliptic operator.

Exactly as in the Riemannian situation, any spinor defines an associated one-form, the so-called *Dirac current*. In Lorentzian geometry, an additional factor i appears in the definition, basically to balance the effect of the aforementioned differences. As in the Riemannian case, elementary calculations show that the Dirac current of a parallel spinor is a parallel spinor field. While in Riemannian geometry, the Dirac current of a

real Killing spinor is a Killing vector field, in the Lorentzian case the same is true for *imaginary* Killing spinors. In stark contrast to the Riemannian situation, in Lorentzian geometry the Dirac current is always non-trivial for a non-vanishing vector field. The Dirac current of any eigenspinor of a twisted Dirac operator is always divergence-free, and $(\cdot,\cdot)$ can be used to define a conserved charge.

Many properties of spacetimes carrying special spinor fields can be read off from their Dirac current. E.g., as shown by Ehlers and Kundt [46], a four-dimensional Lorentzian spin manifold with a parallel spinor is locally isometric to a pp-wave (in this case, the Dirac current is a parellel null vector field).

An important application of spinors in Lorentzian geometry is the commented proof of the positive mass theorem due to the seminal ideas by Witten [115], made rigorous by Parker and Taubes [90] and others (see the independent work by Reula [97] as well as [66] and references therein). In fact, the spacetime viewpoint is necessary here. So, we will revisit the approach in the previous section, and focus in the positiveness of the energy (which, as commented above, could be also proven by using Schoen and Yau techniques). The starting point is a Cauchy hypersurface $(\Sigma, g|_\Sigma)$ that is asymptotically flat in the sense explained around formula (4), including the bounds for the second fundamental form. In this case, the expression of m in (5) is taken as the definition of the energy E, and the momenta P_l are defined by:

$$P_i = \frac{1}{8\pi} \lim_{r\to\infty} \int_{S_r} \left(\sum_{j=1}^{3} \sigma_{ij} n^j - \sum_{j=1}^{3} \sigma^j_j n_i \right) dA,$$

where $\sum_{j=1}^{3} \sigma^j_j /3$ is the mean curvature of Σ. As we have seen, E is independent of the chosen asymptotic coordinates, and the freedom in the choice of these coordinates yields new momenta P_1', P_2', P_3' which differ in an element of $O(3)$. This allows to construct a vector $V := (E, P_1, P_2, P_3) \in \mathbb{R}^{1,3}$, the *ADM energy-momentum* (a different choice Σ' of hypersurface would yield a new vector V' which would be related to V). The statement of the positive energy theorem is that V is a causal vector and equal to 0 if and only if the spacetime is flat around Σ. Witten's idea was to use a Weitzenböck formula for the spacetime Dirac operator applied to spinors tangent to the hypersurface and extended parallely along normal geodesics in a small normal neighborhood of the Cauchy hypersurface. The dominant energy condition then ensures that the zeroth-order term in the Weitzenböck formula is positive, that yields directly the positiveness of the energy. Moreover, one can find a harmonic spinor approaching in coordinates a parallel spinor on $\mathbb{R}^3$ (thought of as in embedded via the Cauchy hypersurface) such that the limit of the boundary term appearing in the integral form of the Weitzenböck formula is exactly $E - |P|$, thus obtaining $E \geq |P|$ i.e., the energy momentum is a causal vector (a corollary is then the positive mass theorem $E \geq 0$). A central tool to prove existence of these solutions are Green's functions in weighted Sobolev spaces performed in detail by Parker and Taubes. In 1987, Yip [116] showed that the energy-momentum vector has to be even *timelike* (non-lightlike) unless M is flat around the Cauchy hypersurface, also by using spinors techniques. As already pointed out, Eichmair et al. [47] have given a recent proof of the energy-momentum inequality $E \geq |P|$ in the case that the manifold M is not necessarily spin, but that $\dim(M) \leq 7$. This is obtained by using Schoen and Yau techniques but, remarkably,

a new difficulty appears, as minimal surfaces are now replaced by marginally outer trapped hypersurfaces, which do not come from a variational characterization.

Another fundamental concept in geometry connected to spinors is *holonomy*. That notion can be defined on any bundle with a connection, denoting, at a point p, the group of diffeomorphisms of the fiber over p which are parallel transports along curves starting and ending at p. If applied to a semi-Riemannian manifold of signature (m, n) with its Levi- Civita connection, it is a restriction of the standard representation of $SO(m, n)$ to a subgroup. It is easy to see that in a connected manifold, the equivalence class of the holonomy representation does not depend on the point p. The corresponding infinitesimal notion (taking the Lie algebra of the holonomy group) is called the *holonomy algebra*.

Now, in the Riemannian case, we have Berger's list: a simply-connected Riemannian manifold is either locally symmetric or can be decomposed as a Riemannian product each k-dimensional factor of which has the holonomy $SO(k)$, $U(k/2)$, $SU(k/2)$, $Sp(k/4) \cdot SP(1)$, $Sp(k/4)$, G_2 (in which case $k = 7$) or $Spin(7)$ (in which case $k = 8$). The Lorentzian case is a bit more involved and has remained open until recently. One of the difficulties compared to the Riemannian case is the difference between decomposability and reducibility pointed out in Sect. 2. In fact, classical de Rham Riemannian decomposition relies on the fact that the orthogonal complement $A^{\perp}$ of an invariant subspace A must be not only invariant too, but also a complement of A. When the latter property holds in the semi-Riemannian case, an analogous de Rham-Wu decomposition is obtained, but this is not the case when A is degenerate—something that can occur in the Lorentzian case. So, the elementary building blocks of the Riemannian classification, irreducible subspaces, have therefore to be complemented by new, properly Lorentzian, building blocks, the indecomposable but non-irreducible subspaces. Such an m-dimensional subspace contains an invariant light-like subspace N, and its holonomy algebra is contained in $(\mathbb{R} \oplus so(m-2)) \ltimes \mathbb{R}^{m-2}$, thus a central part of the classification is done by the $so(m)$-projections of the possible holonomy representation, the so-called *screen holonomy* acting on the associated invariant codimension-2 subbundle of τ_M, the *screen bundle* given by $N^{\perp}/N$ which carries a well-defined Riemannian metric. Thomas Leister in his PhD thesis in 2005 (published in [73]) showed that the screen holonomy is always the holonomy algebra of a Riemannian manifold (and, as remarked by Anton Galaev, this is an exceptional feature of Lorentzian geometry not present in higher signatures). In that manner, he was able to solve some of the remaining problems in the Lorentzian classification by means of the corresponding Riemannian techniques, and finally obtained the full classification. Galaev [54] gave then analytic examples for every holonomy representation of Leistner's list. Still, a missing piece were examples of *globally hyperbolic* manifolds *with complete Cauchy hypersurfaces* with the given holonomy representations. This was done for manifolds with parallel spinors (for which case the above classification yields groups $G \ltimes \mathbb{R}^n$ for G being a product of $SU(p), Sp(q), G_2$ or $Spin(7)$) in an article of Helga Baum and Olaf Müller [14], via a cylinder construction analogous to one by Bär, Gauduchon and Moroianu [8], building a parallel spinor from a so-called Codazzi spinor, and another construction relating Codazzi spinors to imaginary Killing spinors (whose importance in geometry is explained in the next paragraph). In 2013, Helga Baum and Thomas Leistner then solved the analytic initial value problem for parallel spinors [16].

A *Killing spinor* is a spinor ψ such that there is a constant $b \in \mathbb{C}$ with $\nabla_X \psi = bX \cdot \psi$ for all vectors X. As shown by Friedrich in [53], Killing spinors can serve as landmarks where spectral estimates get sharp, in the following sense: If M is compact and the scalar curvature is bounded from below by a positive constant s_0, then for all eigenvalues a of the Dirac operator we have $a^2 \geq \frac{1}{4}\frac{n}{n-1}s_0$, and equality in this estimate implies that the corresponding eigenspinor is a Killing spinor. One can consider the modified connection $\nabla_b := \nabla - b\mathbf{1}$ to conclude that a Killing spinor never vanishes. An elementary calculation shows that $(\mathrm{Ric}(X) - 4b^2(n-1)X) \cdot \psi = 0$, and that implies (by nonvanishing of ψ) that the image $\mathrm{Ric} - 4b^2(n-1)\mathbf{1}_{TM}$ is contained in the null cone. Taking the trace once more, one sees that the scalar curvature equals $4n(n-1)b^2$, in particular, b is either real or purely imaginary. A Killing spinor is called real resp. imaginary if $b \in \mathbb{R}$ resp. $b \in i\mathbb{R}$. Over several years, different people aimed at a full classification of Killing spinors. Christian Bär [7] finally came up with a cone construction which associated to each Killing spinor on a manifold M a parallel spinor on the Riemannian cone over M. As the existence of parallel spinors leads to special holonomy, Bär obtained a classification via the classification of Riemannian holonomies. Imaginary Killing spinors were classified by Helga Baum in [12, 13] in a completely different way: Let (M, g) be a complete connected spin manifold. It carries an $i \cdot a$-Killing spinor iff it is a warped product $\mathbb{R} \times_{e^{-4at}} N$ for a complete connected spin manifold with a non-zero parallel spinor field. The idea is to show that the manifold is foliated by level sets of the norm t of the spinor field. Christoph Bohle [25] examined real Killing spinors on Lorentzian manifolds, relating them also to warped products. Felipe Leitner [74], finally, considered imaginary Killing spinors on Lorentzian manifolds. Their Dirac current is easily seen to be causal, and when it is null, then the manifold is Einstein.

9 Some Further Topics and a Double Invitation

In this article, as announced, we intend to invite experts from other branches of mathematics, especially in Riemannian geometry and global analysis, in two respects: Firstly, we invite *users* of Lorentzian geometry. We hope that the article made clear that Lorentzian geometry can be extremely useful not only in physics, but also in mathematical contexts. One famous example is the aforementioned solution of the Riemannian Yamabe problem via Lorentzian techniques. To a large extent, this potential of Lorentzian geometry remains unexplored up to now.

Secondly, we want to invite *providers*. The open topics in Lorentzian geometry do need support from other branches of mathematics. In the following, we list some important open questions in Global Lorentzian Geometry (without any claim of completeness). In order to do so, it is convenient to distinguish between those that arise directly in Mathematical Relativity and those that are mathematically natural, independent of physical motivations. Along this paper we have emphasized some of the first type. But recall that the questions on Lorentzian manifolds inspired only in reasons of mathematical naturalness and beauty, are interesting in their own right and, sooner or later, will have applications to General Relativity or other parts of Mathematical Physics—recall that General Relativity is one of the two fundamental physical theories, and Quantum Theory the other one. For decades, the attempt to unify

both theories has been a physical challenge and a permanent source of mathematical inspiration.

Along this article some open questions in Mathematical Relativity has appeared more or less explicitly, such as: (a) Cosmic Censorship Conjecture (weak and strong), including full Penrose inequality, (b) Cauchy problem (blow up criteria, global regularity for large data ...), or (c) definitively satisfactory definition of singularities, including both singularity theorems (which involve divergences on curvature and not merely incompleteness) and a precise description of the *boundary* of the spacetime. Of course, there are many other relevant questions in Mathematical Relativity (see [37]). We would like to point out here the interest attracted by the questions of stability. Christodoulou and Klainerman [36] proved the non-linear stability of Lorentz-Minkowski spacetime $\mathbb{L}^4$ as a solution of Einstein equation. This means that a small perturbation of the initial conditions for $\mathbb{L}^4$ yields a spacetime with properties close to $\mathbb{L}^4$ (and, for example, not to a spacetime with singularities). In spite of the simplicity of this idea, the proof is extremely difficult-recall that [36] is a 500 pages book. The result is a landmark in Mathematical Relativity, and opens the study of the stability under weaker falloff hypotheses of the initial data or the stability of other spacetimes, as those with constant curvature, or of the Einstein equation coupled to other field theories.

Finally, let us point out some more purely mathematical questions, some of them suggested above, but only tangentially.

(1) Classification of submanifolds with natural geometric properties (constant mean curvature, umbilic, etc.) in spaceforms and other physical or mathematically relevant spacetimes; notice that some of these questions had motivations from the viewpoint of the initial value problem and were commented in Sect. 5, but such problems evolve further, independent of physical motivations.
(2) Critical curves for indefinite functionals on Lorentzian manifolds: even though the role of geodesics in General Relativity gives a general support for this, the infinite-dimensional variational mathematical approach for geodesics, including spacelike ones, has independent interest, see the seminal works by Benci, Fortunato and Giannoni [19], the book [81] of the review [28]; we emphasize that even a simple question as if any compact Lorentzian manifold must admit a closed geodesic remains open.
(3) Curvature: curvature bounds groups have been stressed above, but there are many other questions related with curvature operators, e.g., those starting at the Osserman problem, solved a decade ago, see [58].
(4) Classification of Lorentzian spaceforms: such a topic has a deep importance and tradition in Geometry, we recommend the recent revision of a paper by Mess in [5] as an example of this exciting problem.
(5) Links between Lorentzian and Finslerian geometries at different levels are being developed fast in the last years, see [30, 31, 51] as a sampler.

Acknowledgements The second-named author is partially supported by the Grants MTM2010–18099 (MICINN) and P09-FQM-4496 (J. Andalucía) with FEDER funds.

References

1. Alías, L.J., Romero, A., Sánchez, M.: Spacelike hypersurfaces of constant mean curvature in certain spacetimes. Nonlinear Anal. **30**, 655–661 (1997)
2. Alías, L.J., Chaves, R.M.B., Mira, P.: Björling problem for maximal surfaces in Lorentz-Minkowski space. Math. Proc. Camb. Philos. Soc. **134**(2), 289–316 (2003)
3. Alías, L.J., Impera, D., Rigoli, M.: Hypersurfaces of constant higher order mean curvature in warped products. Trans. Am. Math. Soc. **365**(2), 591–621 (2013)
4. Alías, L.J., Montiel, S.: Uniqueness of spacelike hypersurfaces with constant mean curvature in generalized Robertson-Walker spacetimes. In: Differential Geometry, Valencia, 2001, pp. 59–69. World Sci., River Edge (2002)
5. Andersson, L., Barbot, T., Benedetti, R., Bonsante, F., Goldman, W.M., Labourie, F., Scannell, K.P., Schlenker, J.-M.: Notes on a paper of Mess [Lorentz spacetimes of constant curvature, Geom. Dedicata **126**, 3–45 (2007)]. Geom. Dedicata **126**, 47–70 (2007)
6. Andersson, L., Howard, R.: Comparison and rigidity theorems in semi-Riemannian geometry. Commun. Anal. Geom. **6**, 819–877 (1998)
7. Bär, C.: Real Killing spinors and holonomy. Commun. Math. Phys. **154**(3), 509–521 (1993)
8. Bär, C., Moroianu, A., Gauduchon, P.: Generalized cylinders in semi-Riemannian and spin geometry. Math. Z. **249**, 545–580 (2005)
9. Bartnik, R.: Existence of maximal surfaces in asymptotically flat spacetimes. Commun. Math. Phys. **94**, 155–175 (1984)
10. Bartnik, R., Chrusciel, P.T., Murchadha, N.O.: On maximal surfaces in asymptotically flat spacetimes. Commun. Math. Phys. **130**, 95–109 (1990)
11. Bartnik, R., Isenberg, J.: The constraint equations. In: The Einstein Equations and the Large Scale Behavior of Gravitational Fields, pp. 1–38. Birkhäuser, Berlin (2004)
12. Baum, H.: Riemannian manifolds with imaginary Killing spinors. Ann. Glob. Anal. Geom. **7**(2), 141–154 (1989)
13. Baum, H.: Complete Riemannian manifolds with imaginary Killing spinors. Ann. Glob. Anal. Geom. **7**(3), 205–226 (1989)
14. Baum, H., Müller, O.: Codazzi spinors and globally hyperbolic Lorentzian manifolds with special holonomy. Math. Z. **258**, 185–211 (2008)
15. Baum, H.: Eichfeldtheorie. Springer, Berlin (2009)
16. Baum, H., Leistner T.: in preparation
17. Beem, J.K., Ehrlich, P.E., Easley, K.L.: Global Lorentzian Geometry. Monographs Textbooks Pure Appl. Math., vol. 202. Marcel Dekker, New York (1996)
18. Benavides Navarro, J.J., Minguzzi, E.: Global hyperbolicity is stable in the interval topology. J. Math. Phys. **52**, 112504 (2011)
19. Benci, V., Fortunato, D., Giannoni, F.: On the existence of multiple geodesics in static space–times. Ann. Inst. Henri Poincaré, Anal. Non Linéaire **8**, 79–102 (1991)
20. Berline, N., Getzler, E., Vergne, M.: Heat Kernels and Dirac Operators. Springer, Berlin (1992)
21. Bernal, A.N., Sánchez, M.: On smooth Cauchy hypersurfaces and Geroch's splitting theorem. Commun. Math. Phys. **243**, 461–470 (2003)
22. Bernal, A.N., Sánchez, M.: Smoothness of time functions and the metric splitting of globally hyperbolic spacetimes. Commun. Math. Phys. **257**, 43–50 (2005)
23. Bernal, A.N., Sánchez, M.: Further results on the smoothability of Cauchy hypersurfaces and Cauchy time functions. Lett. Math. Phys. **77**, 183–197 (2006)
24. Bernal, A.N., Sánchez, M.: Globally hyperbolic spacetimes can be defined as "causal" instead of "strongly causal". Class. Quantum Gravity **24**, 745–750 (2007)
25. Bohle, C.: Killing spinors on Lorentzian manifolds. J. Geom. Phys. **45**, 285–308 (2003)
26. Bray, H.: Proof of the Riemannian Penrose inequality using the positive mass theorem. J. Differ. Geom. **59**, 177–267 (2001)
27. Bray, H.: Black holes, geometric flows, and the penrose inequality in general relativity. Not. Am. Math. Soc. **49**, 1372–1381 (2003)
28. Candela, A.M., Sánchez, M.: Geodesics in semi-Riemannian manifolds: geometric properties and variational tools. In: Recent Developments in Pseudo-Riemannian Geometry. ESI Lect. Math. Phys. pp. 359–418. Eur. Math. Soc., Zurich (2008)
29. Caponio, E., Germinario, A.V., Sánchez, M.: Convex regions of stationary spacetimes and Randers spaces. Applications to lensing and asymptotic flatness (2011). arXiv:1112.3892

30. Caponio, E., Javaloyes, M.A., Masiello, A.: On the energy functional on Finsler manifolds and applications to stationary spacetimes. Math. Ann. **351**, 365–392 (2011)
31. Caponio, E., Javaloyes, M.A., Sánchez, M.: On the interplay between Lorentzian causality and Finsler metrics of Randers type. Rev. Mat. Iberoam. **27**, 919–952 (2011)
32. Carrasco, A., Mars, M.: A counterexample to a recent version of the Penrose conjecture. Class. Quantum Gravity **27**(6), 062001 (2010)
33. Chaves, R.M.B., Dussan, M.P., Magid, M.: Björling problem for timelike surfaces in the Lorentz-Minkowski space. J. Math. Anal. Appl. **377**(2), 481–494 (2011)
34. Cheng, S.-Y., Yau, S.-T.: Maximal space-like hypersurfaces in the Lorentz-Minkowski spaces. Ann. Math. **104**, 407–419 (1976)
35. Choquet-Bruhat, Y., Geroch, R.: Global aspects of the Cauchy problem in General Relativity. Commun. Math. Phys. **14**, 329–335 (1969)
36. Christodoulou, D., Klainermann, S.: On the Global Nonlinear Stability of Minkowski Space. Princeton University Press, Princeton (1995)
37. Chruściel, P., Galloway, G.J., Pollack, D.: Mathematical general relativity: a sampler. Bull. Am. Math. Soc. (N.S.) **47**(4), 567–638 (2010)
38. Chruściel, P.T., Grant, J., Minguzzi, E.: On differentiability of volume time functions. Preprint (2013)
39. Chrusciel, P.T., Isenberg, J., Pollack, D.: Initial data engineering. Commun. Math. Phys. **257**, 29–42 (2005)
40. Chruściel, P.T., Shatah, J.: Global existence of solutions of the Yang-Mills equations on globally hyperbolic four dimensional Lorentzian manifolds. Asian J. Math. **1**, 530–548 (1997)
41. Corvino, J.: Scalar curvature deformation and a gluing construction for the Einstein constraint equations. Commun. Math. Phys. **214**, 137–189 (2000)
42. Corvino, J., Schoen, R.: On the asymptotics for the vacuum Einstein constraint equations. J. Differ. Geom. **73**, 185–207 (2006)
43. Dafermos, M.: Spherically symmetric spacetimes with a trapped surface. Class. Quantum Gravity **22**, 2221–2232 (2005)
44. Ecker, K.: On mean curvature flow of spacelike hypersurfaces in asymptotically flat spacetimes. J. Aust. Math. Soc. A **55**(1), 41–59 (1993)
45. Ecker, K., Huisken, G.: Parabolic methods for the construction of spacelike slices of prescribed mean curvature in cosmological spacetimes. Commun. Math. Phys. **135**(3), 595–613 (1991)
46. Ehlers, J., Kundt, W.: Exact solutions of the gravitational field equation. In: Witten, L. (ed.) Gravitation, an Introduction to Current Research, pp. 49–101. Wiley, New York (1962)
47. Eichmair, M., Huang, L.-H., Lee, D.A., Schoen, R.: The spacetime positive mass theorem in dimensions less than eight. arXiv:1110.2087
48. Ellis, G.F.R., Hawking, S.W.: The Large Scale Structure of Space-Time. Cambridge Monographs on Mathematical Physics, vol. 1. Cambridge University Press, London (1973)
49. Fathi, A., Siconolfi, A.: On smooth time functions. Math. Proc. Camb. Philos. Soc. **152**(2), 303–339 (2012)
50. Flores, J.L., Herrera, J., Sanchez, M.: On the final definition of the causal boundary and its relation with the conformal boundary. Adv. Theor. Math. Phys. **15**(4), 991–1058 (2011)
51. Flores, J.L., Herrera, J., Sánchez, M.: Gromov, Cauchy and causal boundaries for Riemannian, Finslerian and Lorentzian manifolds. Mem. Am. Math. Soc. **226**, 1064 (2013)
52. Frauendiener, J.: Conformal infinity. Living Rev. **1** (2004). http://relativity.livingreviews.org/Articles/lrr-2004-1/index.html
53. Friedrich, T.: Der erste Eigenwert des Dirac-Operators einer kompakten Riemannschen Mannigfaltigkeit nichtnegativer Skalarkrümmung. Math. Nachr. **97**, 117–146 (1980)
54. Galaev, A.: Metrics that realize all Lorentzian holonomy algebras. Int. J. Geom. Methods Mod. Phys. **3**(5–6), 1025–1045 (2006)
55. Galloway, G.J., Senovilla, J.M.M.: Singularity theorems based on trapped submanifolds of arbitrary co-dimension. Class. Quantum Gravity **27**(15), 152002 (2010)
56. García-Parrado, A., Sánchez, M.: Further properties of causal relationship: causal structure stability, new criteria for isocausality and counterexamples. Class. Quantum Gravity **22**, 4589–4619 (2005)
57. García-Parrado, A., Senovilla, J.M.M.: Causal relationship: a new tool for the causal characterization of Lorentzian manifolds. Class. Quantum Gravity **20**, 625–664 (2003)
58. García-Río, E., Kupeli, D.N., Vázquez-Lorenzo, R.: Osserman Manifolds in Semi-Riemannian Geometry. Lecture Notes in Mathematics, vol. 1777. Springer, Berlin (2002)

59. Gerhardt, C.: H-surfaces in Lorentzian manifolds. Commun. Math. Phys. **89**, 523–553 (1983)
60. Gerhardt, C.: Hypersurfaces of prescribed mean curvature in Lorentzian manifolds. Math. Z. **235**(1), 83–97 (2000)
61. Gerhardt, C.: On the CMC foliation of future ends of a spacetime. Pac. J. Math. **226**, 297–308 (2006)
62. Gerhardt, C.: Curvature Problems. Series in Geometry and Topology, vol. 39. International Press, Somerville (2006)
63. Geroch, R.: What is a singularity in general relativity? Ann. Phys. **48**, 526–540 (1968)
64. Geroch, R.: Domain of dependence. J. Math. Phys. **11**, 437–449 (1970)
65. Helfer, A.: Conjugate points on spacelike geodesics or pseudo-selfadjoint Morse-Sturm-Liouville systems. Pac. J. Math. **164**(2), 321–350 (1994)
66. Herzlich, M.: The positive mass theorem for black holes revisited. J. Geom. Phys. **26**, 97–111 (1998)
67. Huisken, G., Ilmanen, T.: The Inverse mean curvature flow and the Riemannian Penrose inequality. J. Differ. Geom. **59**, 353–437 (2001)
68. Isenberg, J., Mazzeo, R., Pollack, D.: Gluing and wormholes for the Einstein constraint equations. Commun. Math. Phys. **231**, 529–568 (2002)
69. Kovner, I.: Fermat principle in gravitational fields. Astrophys. J. **351**, 114–120 (1990)
70. Krasnikov, S.: No time machines in classical general relativity. Class. Quantum Gravity **19**, 4109 (2002)
71. Lawson, H.B., Michelson, M.L.: Spin Geometry. Princeton University Press, Princeton (1989)
72. Lee, J.M., Parker, T.: The Yamabe problem. Bull. Am. Math. Soc. **17** (1987)
73. Leistner, T.: On the classification of Lorentzian holonomy groups. J. Differ. Geom. **76**, 423–484 (2007)
74. Leitner, F.: Imaginary Killing spinors in Lorentzian geometry. J. Math. Phys. **44**, 4795 (2003)
75. Lerner, D.E.: The space of Lorentz metrics. Commun. Math. Phys. **32**, 19–38 (1973)
76. Lichnerowicz, A.: L'integration des équations de la gravitation relativiste et le problème des n corps. J. Math. Pures Appl. **23**, 37–63 (1944)
77. Manchak, J.B.: No no-go: a remark on time machines. Preprint (2012)
78. Mars, M.: Present status of the Penrose inequality. Class. Quantum Gravity **26**(19), 193001 (2009)
79. Marsden, J.E., Tipler, F.J.: Maximal hypersurfaces and foliations of constant mean curvature in general relativity. Phys. Rep. **66**(3), 109–139 (1980)
80. Marsden, J.E.: On completeness of homogeneous pseudo-Riemannian manifolds. Indiana Univ. Math. J. **22**, 1065–1066 (1972/1973)
81. Masiello, A.: Variational Methods in Lorentzian Geometry. Longman Sc. Tech, Harlow (1994)
82. Minguzzi, E., Sánchez, M.: Connecting solutions of the Lorentz force equation do exist. Commun. Math. Phys. **264**(2), 349–370 (2006)
83. Minguzzi, E., Sánchez, M.: The causal hierarchy of spacetimes. In: Recent Developments in Pseudo-Riemannian Geometry. ESI Lect. Math. Phys., pp. 299–358. Eur. Math. Soc., Zürich (2008)
84. Montiel, S.: An integral inequality for compact spacelike hypersurfaces in de Sitter space and applications to the case of constant mean curvature. Indiana Univ. Math. J. **37**(4), 909–917 (1988)
85. Müller, O.: The Cauchy problem of Lorentzian minimal surfaces in globally hyperbolic manifolds. Ann. Glob. Anal. Geom. **32**(1), 67–85 (2007)
86. Müller, O.: Asymptotic flexibility of globally hyperbolic manifolds. C. R. Math. Acad. Sci. Paris **350**(7–8), 421–423 (2012)
87. Müller, O.: Special temporal functions on globally hyperbolic manifolds. Lett. Math. Phys. **103**(3), 285–297 (2013)
88. Müller, O., Sánchez, M.: Lorentzian manifolds isometrically embeddable in L^N. Trans. Am. Math. Soc. **363**(10), 5367–5379 (2011)
89. O'Neill, B.: Semi-Riemannian Geometry with Applications to Relativity. Academic Press, New York (1983)
90. Parker, T., Taubes, C.H.: On Witten's proof of the positive energy theorem. Commun. Math. Phys. **84**(2), 223–238 (1982)
91. Penrose, R.: Gravitational Collapse and Space-Time Singularities. Phys. Rev. Lett. **14**, 57–59 (1965)
92. Perlick, V.: On Fermat's principle in general relativity. I. The general case. Class. Quantum Gravity **7**, 1319–1331 (1990)
93. Perlick, V.: On Fermat's principle in general relativity. II. The conformally stationary case. Class. Quantum Gravity **7**, 1849–1867 (1990)
94. Perlick, V.: Gravitational lensing from a spacetime perspective. Living Rev. Relat. **7** (2004). http://relativity.livingreviews.org/Articles/lrr-2004-9/

95. Piccione, P., Tausk, D.V.: On the distribution of conjugate points along semi-Riemannian geodesics. Commun. Anal. Geom. **11**, 33–48 (2003)
96. Pigola, S., Rigoli, M., Setti, A.G.: Maximum principles on Riemannian manifolds and applications. Mem. Am. Math. Soc. **174**, 822 (2005)
97. Reula, O.A.: Existence Theorem for solutions of Witten's equation and non-negativity of the total mass. J. Math. Phys. **23**(5) (1982)
98. Romero, A., Rubio, R.M., Salamanca, J.J.: Uniqueness of complete maximal hypersurfaces in spatially parabolic generalized Robertson-Walker spacetimes. Class. Quantum Gravity **30**(11), 115007 (2013)
99. Romero, A., Sánchez, M.: New properties and examples of incomplete Lorentzian tori. J. Math. Phys. **35**, 1992–1997 (1994)
100. Sachs, R.K., Wu, H.: General relativity and cosmology. Bull. Am. Math. Soc. **83**, 1101–1164 (1977)
101. Sánchez, M.: Causal hierarchy of spacetimes, temporal functions and smoothness of Geroch's splitting. A revision. Mat. Contemp. **29**, 127–155 (2005)
102. Sánchez, M.: Cauchy hypersurfaces and global Lorentzian geometry. In: Proc. XIV Fall Workshop Geom. Phys., Bilbao Spain, September 14-16 2005. Publ. RSME, vol. 20, pp. 2–22 (2006)
103. Sánchez, M.: A note on stability and Cauchy time functions. Preprint. arXiv:1304.5797
104. Schmidt, B.G.: A new definition of singular points in general relativity. Gen. Relativ. Gravit. **1**(3), 269–280 (1970/1971)
105. Schoen, R., Yau, S.-T.: Proof of the positive mass theorem I. Commun. Math. Phys. **65**, 45–76 (1979)
106. Schoen, R., Yau, S.-T.: Proof of the positive mass theorem II. Commun. Math. Phys. **79**, 1457–1459 (1981)
107. Senovilla, J.M.M.: Trapped surfaces, horizons and exact solutions in higher dimensions. Class. Quantum Gravity **19**, L113 (2002)
108. Senovilla, J.M.M.: New class of inhomogeneous cosmological perfect-fluid solutions without big-bang singularity. Phys. Rev. Lett. **64**(19), 2219–2221 (1990)
109. Senovilla, J.M.M.: A singularity theorem based on spatial averages. In: Dadhich, N., Joshi, P., Roy, P. (eds.) The Raychaudhuri Equation and Its Role in Modern Cosmology, vol. 69, pp. 31–47 (2007)
110. Stumbles, S.M.: Hypersurfaces of constant mean extrinsic curvature. Ann. Phys. **133**(1), 28–56 (1981)
111. Szabados, L.: Quasi-local energy-momentum and angular momentum in GR: a review article. Living Rev. **4** (2004). http://relativity.livingreviews.org/Articles/lrr-2004-4/index.html
112. Treibergs, A.E.: Entire spacelike hypersurfaces of constant mean curvature in Minkowski space. Invent. Math. **66**(1), 39–56 (1982)
113. Uhlenbeck, K.: A Morse theory for geodesics on a Lorentz manifold. Topology **14**, 69–90 (1975)
114. Wald, R.M.: General Relativity. Univ. Chicago Press, Chicago (1984)
115. Witten, E.: A new proof of the positive energy theorem. Commun. Math. Phys. **80**, 381–402 (1981)
116. Yip, P.F.: A strictly-positive mass theorem. Commun. Math. Phys. **108**(4), 653–665 (1987)

Olaf Müller obtained his PhD at the Max-Planck Institute for Mathematics in the Sciences, Leipzig, in 2004. Since then, he has worked at the Humboldt University Berlin in the group of Helga Baum, at the Universidad Nacional Autónoma de México (UNAM) and at the University of Regensburg in the group of Felix Finster. His research focuses on differential geometry and global analysis, in particular the geometry of globally hyperbolic manifolds.

Miguel Sánchez is a mathematician and physicist at the Department of Geometry and Topology of the University of Granada. He obtained his PhD in Mathematics in 1994 and, since then, has worked on several topics in Differential Geometry, Global Analysis on Manifolds and Mathematical Physics. He has supervised four PhD thesis, and his publications include more than fifty research articles, as well as a couple of text books, one of them about Lorentzian Geometry.

Jahresber Dtsch Math-Ver (2014) 115:185–206
DOI 10.1365/s13291-013-0071-5

HISTORICAL ARTICLE

Hilberts „*Grundlagen der Geometrie*“ und ihre Stellung in der Geschichte der Grundlagendiskussion

Ulrich Felgner

Online publiziert: 30. Oktober 2013

Zusammenfassung In seinen „*Grundlagen der Geometrie*“ hat Hilbert die Bemühungen um eine Grundlegung der Geometrie zu einem überzeugenden Abschluß gebracht. Um diese Leistung würdigen zu können, müssen wir auf die lange Geschichte der Grundlagendiskussion – von der Antike bis zur Gegenwart – jedenfalls in großen Zügen eingehen. Wir werden insbesondere über das Problem, wie die geometrischen Grundbegriffe einzuführen sind (Euklid, Heron, Descartes, Pascal, Hobbes, Tschirnhaus et al.), und über die verschiedenen Entwürfe einer axiomatischen Grundlegung (Aristoteles, Euklid, Tschirnhaus et al.) berichten und danach das von Hilbert aufgestellte Axiomensystem besprechen. Ähnlich wie Dedekind 1888 den Bereich der natürlichen Zahlen (bis auf Isomorphie) als *minimales* Modell eines bestimmten Axiomensystems charakterisieren konnte, gelang es Hilbert, die euklidische Geometrie als *maximales* Modell seines Axiomensystems zu charakterisieren.

Schlüsselwörter Euklidische Geometrie · Axiomatik · Strukturalismus · Implizite Definitionen · das Problem der Anschauung

Mathematics Subject Classification (2000) 00A30 · 01A20 · 01A55 · 03A05 · 51-03

Einleitung: *Die Publikation von* Hilberts „*Grundlagen der Geometrie*“, *1899*

Hilbert beschäftigt sich in seinem Buch mit den *Grundlagen* der Geometrie. Er will „*eine kritische Untersuchung der Prinzipien der Geometrie*“ vornehmen, wie er selber in seinem Schlußwort hervorhebt. Es geht in dem Buch also nicht darum, die Geometrie *von Grund auf* zu entwickeln, sondern nur darum, *den Grund zu legen*,

U. Felgner (✉)
Mathematisches Institut, Universität Tübingen, Tübingen, Deutschland
e-mail: felgner@uni-tuebingen.de

auf dem das Gebäude der Geometrie errichtet werden kann. Das Buch will also kein Lehrbuch der Geometrie von der Art der *Elemente* Euklids, des *Euclides restitutus* von Alphonso Borelli, oder Legendres *Éléments de Géométrie* oder ähnlichen Werken sein. Es will vielmehr ein Buch sein, das die *logischen* und *ontologischen* Probleme, die die Geometrie aufwirft, lösen möchte.

Die *logischen Probleme* betreffen den axiomatischen Aufbau der Geometrie. Was noch niemandem vor Hilbert gelungen war, wird hier erstmals ausgeführt: es wird ein *vollständiges* Axiomensystem für die euklidische Geometrie angegeben und darüber hinaus die Kategorizität (Monomorphie) und die relative Widerspruchsfreiheit dieses Axiomensystems nachgewiesen. Von zahlreichen Axiomen wird ferner die Tragweite geklärt und die Unabhängigkeit bewiesen.

Die *ontologischen Probleme* betreffen die Natur und den Seinsstatus der geometrischen Gegenstände. Was sind „Punkte", „gerade Linien" etc.? Da sie ausdehnungslos bzw. breitenlos sein sollen, können sie nicht sinnlich wahrnehmbar sein und können auch nicht wirklich vorgestellt werden. Man darf mit Recht fragen, ob es sie wirklich gibt, oder in welchem Sinne es sie gibt. Gibt es Objekte, die *als* „Punkte" oder *als* „Linien" etc. angesehen werden können und die den üblichen Axiomen der Geometrie genügen?

Auch auf diese Frage, die von Mathematikern und Philosophen seit der Antike kontrovers diskutiert wurde, hat Hilbert eine überzeugende Antwort gegeben und zwar im Sinne des *Strukturalismus*. [Darüber werden wir in Abschn. 7 noch ausführlich sprechen.] Die geometrischen Objekte müssen nach Hilberts Auffassung kein raum-zeitliches Substrat haben; sie verdanken ihre Existenz lediglich einem formalen Axiomensystem und haben die geforderten Eigenschaften nicht *per se*, sondern nur im Sinne eines Modells des Axiomensystems.

In Hilberts Antwort kommt, wie Otto Blumenthal in seiner *Lebensgeschichte Hilberts* schrieb ([1], S. 403), eine *„radikale Abstraktion von der Anschauung"* zum Ausdruck und ihr verdanken die *Grundlagen der Geometrie* ihren *„gewaltigen Erfolg"*. Der Erfolg wurde gekrönt durch den Lobatschewski-Preis, den Hilbert von der Universität Kasan im Jahre 1904 verliehen bekam.

In seinem Werk *Grundlagen der Geometrie* [11] aus dem Jahre 1899 hat Hilbert *„auf viele der hierher gehörigen Fragen… zum ersten Male eine befriedigende, ja endgültige Antwort"* gegeben (Friedrich Engel in seinem Referat im Jahrbuch über die Fortschritte der Mathematik, Band 30, Berlin 1901, S. 424–426). Hilbert hat die Bemühungen um eine befriedigende Grundlegung der Geometrie zu einem Abschluß geführt. – Der Abschluß war dennoch nur ein vorläufiger Abschluß, denn die hilbertschen Antworten konnten im Laufe des 20. Jahrhunderts noch vertieft und auch sehr viel klarer und stringenter formuliert werden. Insofern steht das hilbertsche Werk in einer Entwicklung, die von der Antike bis in die Gegenwart reicht.

Um die Stellung der hilbertschen *Grundlagen der Geometrie* in der langen Geschichte der Grundlagendiskussion soll es in diesem Essay gehen.[1] Um diese Stel-

[1] Die Stellung der hilbertschen *Grundlagen der Geometrie* in der Geschichte der Geometrie soll hier nicht besprochen werden. Darüber geben die Aufsätze, Essays und Bücher von Freudenthal [7, 8], Hallett-Majer [10], Poincaré [18] und Toepell [24], die wir im Literaturverzeichnis angegeben haben, detaillierte Auskunft.

lung angemessen würdigen zu können, müssen wir sehr weit in die Geschichte der Geometrie zurückblicken, denn sie beginnt in der Antike.

1 Die pythagoräische Konzeption der Geometrie

In der Zeit zwischen Thales und Euklid fand eine tiefgreifende Umgestaltung der Geometrie statt und diese Umgestaltung warf Probleme auf, die es in der älteren Geometrie nicht gegeben hatte und um deren Lösung die Mathematiker bis ins 20. Jahrhundert hinein gerungen haben.

Die geometrischen Gegenstände waren für die Ägypter und Babylonier und auch für die Griechen in vorpythagoräischer Zeit die gezeichneten geraden Linien und ihre Schnittpunkte, die gezeichneten Kreise und die von den gezeichneten Linien eingeschlossenen Flächen, etc. In der Astronomie betrachtete man auch Punkte, Linien und Flächen, die man sich in das Himmelsgewölbe eingezeichnet dachte. Von den gezeichneten und den gedachten Linien beachtete man nicht die Breite; man ging aber nicht soweit anzunehmen, daß sie überhaupt keine Breite hätten. Auch von den Punkten nahm man nicht an, daß sie ausdehnungslos wären (sie wären andernfalls auch gar nicht wahrnehmbar gewesen). Genauso wie die Existenz all dieser ausgedehnten Größen bereits durch den Augenschein gesichert ist, sind auch die geometrischen Aussagen, die sich auf solche ausgedehnten Größen beziehen, Feststellungen, die durch die sinnliche Anschauung verifiziert werden können.

Aber etwa vom ausgehenden 5. Jahrhundert v.u.Z. an entwickelte sich in Griechenland eine Geometrie, die nicht mehr von den sinnlich wahrnehmbaren geometrischen Punkten, Linien und Flächen handelte, sondern nur noch von abstrakten oder idealisierten Objekten. Die Wahrheit einer behaupteten Aussage ließ sich nicht mehr mit den Sinnen feststellen, sondern nur noch mit den „Augen des Geistes“, also mit einem begrifflich geführten Beweis.

Ausgelöst wurde diese Entwicklung durch eine Entdeckung, die (nach dem Zeugnis von Iamblichos: [14], § 88, S. 246–247) Hippasos von Metapont gemacht hatte. Er entdeckte etwa in den Jahren um 425 v.u.Z., daß in einem Quadrat zwei aneinander stoßende Seiten und die anliegende Diagonale zwar ein rechtwinkliges Dreieck bilden, daß aber ihre Längen kein ganzzahliges pythagoräisches Zahlentripel bilden. Dazu bewies er, daß Seite und Diagonale eines Quadrates in keinem ganzzahligen Verhältnis stehen, oder kurz gesagt (in heutiger Sprechweise), daß $\sqrt{2}$ irrational ist. Der Beweis beruht auf der Feststellung, daß im Prozeß der Wechselwegnahme ($\grave{\alpha}\nu\theta\upsilon\phi\alpha\acute{\iota}\rho\varepsilon\sigma\iota\varsigma$) jede beliebig vorgegebene Strecke der Länge nach unterschritten werden kann. Ein gemeinsames Maß von Seite und Diagonale eines Quadrates müßte also kleiner als jede beliebig vorgegebene Strecke sein. Ein gemeinsames Maß kann es also hier nicht geben. Um sicher zu stellen, daß der Prozeß der Wechselwegnahme durchführbar und *ad infinitum* fortgesetzt werden kann, müssen die auftretenden Hilfslinien alle existieren, d.h., man muß sie alle noch unterscheiden können, und das heißt, daß die auftretenden Geraden alle *keine Breite* haben dürfen. Es war in der Geschichte der Geometrie bis dahin noch nie nötig gewesen, die Frage zu stellen, wie breit geometrische Linien sein dürfen, aber jetzt stellte sich diese Frage erstmals und die Antwort lautete, daß sie *überhaupt keine Breite haben dürfen*. Ganz genauso

stellte sich die Frage, wie dick Punkte sein dürfen, und die Antwort lautete, daß sie *keine Ausdehnung haben dürfen*.

Damit ergab sich aber auch als Konsequenz, daß die Frage, ob Seite und Diagonale eines Quadrates kommensurabel sind, offenbar nicht auf der Grundlage sinnlicher Wahrnehmung (Empirie) beantwortet werden kann.

Das entscheidend Neue an der Argumentation von Hippasos war, daß sie sich nicht auf die sinnlich wahrnehmbaren Punkte, Linien, Flächen etc. beziehen konnte, sondern sich auf idealisierte Punkte ohne Ausdehnung, idealisierte Linien ohne Breite etc. beziehen mußte. Von solchen idealisierten Objekten war in der älteren ägyptisch-babylonischen Geometrie und auch bei Thales nie die Rede gewesen. Die griechischen Geometer sprachen aber von nun an fast nur noch von diesen idealisierten Objekten und von ihnen handeln auch die euklidischen *Elemente* [4] (geschrieben um 300 v.u.Z.). Es wird gleich im ersten Buch der *Elemente* definiert, daß Punkte solche Gegenstände sind, *„die keine Teile haben“*, und daß Linien *„breitenlose Längen“* seien etc.

Damit hatte sich der Bereich der Gegenstände, der in der Geometrie untersucht werden soll, grundlegend geändert. Die Gegenstände sollen nicht mehr die sinnlich wahrnehmbaren Punkte, Linien und Flächen sein, sondern von nun an die gedachten, idealisierten Punkte, Linien, Flächen etc. Es handelt sich also um Gegenstände, die zwar „nicht von dieser Welt“ sind, die aber dennoch ihr Fundament in der sinnlichen Anschauung haben.

Damit hatte sich auch der Begriff des Beweises in der Geometrie geändert. Man kann sich nicht mehr auf den Augenschein berufen, sondern muß die Gegenstände begrifflich fassen und die Beweise begrifflich führen.

Dabei entdeckten die griechischen Geometer, daß sich zwar vieles durch sinnliche Wahrnehmung auf einen Blick hin erfassen läßt, daß aber nicht alles, was sich durch Augenscheinnahme feststellen läßt, auch durch logisches Fortschreiten von Begriff zu Begriff nachvollziehen läßt. Die *Geradheit* einer Strecke beispielsweise läßt sich mit den Sinnen sehr leicht wahrnehmen, kann aber mit elementar-geometrischen Begriffen allein nicht beschrieben werden.[2] – Die sinnliche Wahrnehmung kann aber die logische Struktur nicht sichtbar machen und ist zudem in der Regel unsicher.

Proklos bemerkte dazu in seinem Euklid-Kommentar ([19], S. 322, bzw. 389), daß manches *„für den Augenschein klar ist, aber für das wissenschaftliche Denken“* oft nur sehr schwer zu analysieren und zu erklären ist.

Die Bemühungen um *„die logische Analyse unserer räumlichen Anschauung“* (cf. Hilbert in der Einleitung zu seinen *Grundlagen der Geometrie*) führte Euklid zur Entdeckung eines grundlegenden Prinzips, das vorher nicht bemerkt worden war: das *Parallelen-Postulat*.

Man beachte hier, daß eine Geometrie, die sich nur auf Objekte bezieht, die mit den gewöhnlichen Hilfsmitteln Lineal, Winkelhaken und Zirkel (im Bereich des Sinnlich-Wahrnehmbaren) konstruierbar sind, kein Parallelen-Postulat benötigt; die

[2] Vergl. dazu die ausführliche Diskussion der euklidischen Definition des Begriffes der „geraden Linie“ in Thomas L. Heath: *The thirteen books of Euclid's Elements*. Cambridge University Press 1956, vol. I, S. 165–169.

eindeutige Existenz von Parallelen ist hier offenbar gültig,[3] denn die Kanten von Linealen und Winkelhaken sind *gerade*. Wenn man aber auf diese Hilfsmittel verzichtet und nur begrifflich vorgehen will, dann läßt sich die eindeutige Existenz von Parallellinien nicht zeigen und man muß postulieren, daß es durch einen vorgegebenen Punkt höchstens eine gibt.[4]

Andere Prinzipien blieben unerkannt, z.B. Archimedizität und Stetigkeit (cf. Abschn. 6).

Die Änderungen in der Auffassung, was Geometrie ist und wie sie durchzuführen ist, waren tiefgreifend und es ist gerechtfertigt, hier von einem *Paradigmenwechsel* zu sprechen.

Die im vierten Jahrhundert v.u.Z. entstandene neue Geometrie hat Euklid in seinen viel bewunderten *Elementen* zusammengefaßt und dargestellt. Es handelt sich um den Versuch einer Darstellung der Geometrie auf ‘axiomatischer’ Grundlage, die von expliziten Definitionen der Grundbegriffe und einer Liste von Postulaten ausgeht. *In den Postulaten wird gefordert, daß die grundlegenden Konstruktionen, die man bisher (in der vorpythagoräischen Geometrie) mit mechanischen Geräten ausführen konnte* (z.B. Verbinden zweier Punkte durch eine Gerade, Verlängern von Geraden, Parallelverschiebungen von Strecken, Schlagen von Kreisen, etc.) *auch in der neuen Geometrie zur Verfügung stehen sollen*.

Euklid spricht allerdings ganz bewußt von *Postulaten* (αἰτήματα) und nicht von *Axiomen*. Es ist aber dennoch in späteren Jahrhunderten üblich geworden, die euklidischen Postulate *Axiome* (im aristotelischen Sinne) zu nennen, obwohl sie – streng genommen – keine *allgemeingültigen, wahren Aussagen* sind. Dementsprechend wird Euklids Vorgehensweise – nicht ganz korrekt – als axiomatische Methode bezeichnet.[5]

2 Die Gegenüberstellung der euklidischen Axiomatik und der aristotelischen Wissenschaftslehre in der Renaissance

Mit dem Untergang der antiken Welt gerieten auch die *Elemente* Euklids in Vergessenheit und schon bald war in West-Europa nirgendwo mehr ein Exemplar dieses einstmals so bewunderten Werkes vorhanden. Erst vom 12. Jahrhundert an wurden

[3] In einer solchen Geometrie kann unter Verwendung von zwei Linealen jede gerade Strecke verlängert werden, indem man die beiden Lineale aneinander gleitend verschiebt. Durch Verschieben eines Winkelhakens längs eines Lineals kann man Parallelen zeichnen, etc.

[4] Daraus ergibt sich erneut, daß die *Elemente* Euklids nicht von den mechanisch ausführbaren Konstruktionen mit Zirkel und Lineal (κίρκινος, κανών, circini regulaeque) handeln, sondern von den begrifflich gegebenen Kreisen und Strecken (κύκλος, γραμμή). Das wird in der Literatur häufig übersehen.

[5] Das Wort „*Axiom*“ bedeutet im Griechischen „*Grundsatz, Grundwahrheit*“. Nach Aristoteles ist ein *Axiom* eine Aussage, die unmittelbar gewiß ist und deshalb unter den ersten Sätzen einer jeden Theorie auftreten kann. In dem euklidischen Postulaten-System werden jedoch keine grundlegenden Wahrheiten zusammengestellt, wie fast überall behauptet wird, sondern lediglich Vereinbarungen, in denen festgehalten wird, wie mit den *idealisierten* Gegenständen der neuen, theoretischen Geometrie umgegangen werden darf. Die Konstruktionsaufgaben in den *Elementen* Euklids sollen auch keine praktischen Konstruktionen mit Zirkel und Lineal beschreiben, sondern zeigen, daß die jeweiligen Figuren (als ideelle Gebilde, bestehend aus Punkten, Kreisen und Strecken) dem reinen, begrifflichen Denken zur Verfügung stehen.

die *Elemente* Euklids im Abendland in lateinischen Übersetzungen aus dem Arabischen wieder bekannt. Abschriften des griechischen Originals tauchten erst wieder in der zweiten Hälfte des 15. Jahrhunderts auf, als nach der Eroberung von Konstantinopel 1453 durch den türkischen Emir Mehmet II viele Gelehrte des kaiserlichen Hofes und viele Mönche nach Italien flohen und die Handschriften der griechischen Dichter, Philosophen und Wissenschaftler aus ihren Bibliotheken und Klöstern mitnahmen. So kam es, daß im Abendland die Werke von Euklid, zahlreiche Werke von Platon, Proklos und vielen anderen im griechischen Original wieder bekannt wurden. Insbesondere löste das Bekanntwerden vieler platonischer Schriften eine wahre Begeisterung aus. Es begann eine Renaissance des Platonismus, was aber auch zu einem Kampf gegen Aristoteles und die Scholastik führte.

Die griechische Urfassung der *Elemente* Euklids wurde zum ersten Mal 1533 in Basel gedruckt. Der Herausgeber Simon Grynaeus schrieb in seinem Vorwort, daß seiner Meinung nach die axiomatisch aufgebaute Geometrie Euklids der vollkommene Maßstab der wissenschaftlichen Methode sei:

„Geometriae, quae methodi totius absoluta et perfecta formula est."

Damit stellte Grynaeus die führende Rolle, die die aristotelische Wissenschaftslehre bis dahin immer noch spielte, in Frage. Als vollkommener Maßstab der wissenschaftlichen Methode galt damals immer noch die Methode, die Aristoteles in seiner *Zweiten Analytik* (Analytica posteriora) ausgearbeitet hatte. Sie war das unbestrittene Ideal aller Disziplinen.[6] Euklid trat somit in Konkurrenz zur *wissenschaftlichen Methode* des Aristoteles. Es begann eine Diskussion der Vorzüge und der Nachteile der euklidischen Axiomatik einerseits und der aristotelischen Wissenschaftslehre (die auch eine Form der Axiomatik ist) andererseits, die schließlich im 17. und 18. Jahrhundert zu Formulierungen von verschiedenen Neufassungen der axiomatischen Methode führte.

Man begann die *Elemente* Euklids kritischer zu lesen und bemerkte zahlreiche Beweislücken und viele Verstöße gegen die axiomatische Methode. Jacques Peletier beispielsweise fiel in seiner Euklid-Edition ([17], Lyon 1557) auf, daß es Euklid nicht

[6] Über den Aufbau einer wissenschaftlichen Theorie spricht Aristoteles im ersten Buch seiner *Analytica posteriora*. Zunächst muß der *Objektbereich*, der studiert werden soll, festgelegt werden. Es muß gesagt werden, in welcher Hinsicht (d.h. als was) die einzelnen Objekte studiert werden sollen. Danach müssen die *Grundsätze* (*Prinzipien*) des Studiums angegeben werden. Die *allgemeinen Grundsätze* sind die logisch-allgemeingültigen Aussagen, von Aristoteles *„Axiome"* genannt. Die *speziellen Grundsätze* (*Thesen*) zerfallen in zwei Klassen, je nachdem ob sie etwas behaupten oder nur etwas erklären. Die einen sind die *Hypothesen* und die anderen die *Definitionen*. Die Wahrheit der Hypothesen muß sich unmittelbar (oder jedenfalls auf induktivem Wege) aus der sinnlichen Erfahrung ergeben. Die Definitionen können Essential-Definitionen (d.h. Wesens-Definitionen), Kausal-Definitionen oder indirekte Definitionen sein. Aus den Grundsätzen insgesamt gewinnt man durch Beweis (*apodeixis*) die übrigen Aussagen der Theorie. Dabei muß nicht nur gezeigt werden, *daß* (ὅτι) es sich so-und-so verhält, sondern auch, *warum* (διότι) es sich so-und-so verhält.

Auch die Geometrie ist nach Aristoteles nach diesem Muster als Naturwissenschaft (!) aufzubauen. Dabei werden die Objekte (die Linien, Kreise etc.) durch *Aphairesis* (ἀφαίρεσις, Abstraktion) und *Chorismos* (χωρισμός, Verselbständigung) aus den jeweiligen Gattungen sinnlich wahrnehmbarer Gegenstände gewonnen. Aber es ist unklar, wie man hier zur Erkenntnis der „wahren" Hypothesen kommt. Der Aufbau einer „wissenschaftlichen" Geometrie (im aristotelischen Sinne) im gewünschten Umfange ist auf einer solchen Grundlage kaum möglich.

gelungen war, die Gleichheit oder Kongruenz von Winkeln ohne Zuhilfenahme der Anschauung zu definieren. Euklid hatte dazu nur das mechanische Aufeinanderlegen von Winkel angeboten und sich somit auf physikalische Eigenschaften starrer Körper gestützt. Euklid verwandte im Umgang mit Winkeln das Verb $\grave{\varepsilon}\varphi\alpha\varrho\mu\acute{o}\zeta\varepsilon\iota\nu$ (aufeinanderlegen) und das zeigt, daß ihm hier die Elimination der Anschauung nicht gelungen war. Peletier stellte die Frage, warum Euklid in **I**,4 das Aufeinanderlegen von Figuren erlaubt, aber in anderen Beweisen nicht. Wenn es immer erlaubt wäre, dann könnte man viele Beweise erheblich abkürzen.

Bertrand Russell schrieb zum Beweis des Satzes **I**,4 in seinen *Principles of Mathematics* (Cambridge, 1903, S. 404–405):

> *„Indeed Euclid's proof is so bad that he would have done better, to assume this proposition as an axiom.“*

Aber auch ein neues Problem rückte in das Zentrum des Interesses, nämlich wie man die Grundbegriffe der Geometrie (und auch der Arithmetik) einführen soll. Man erkannte immer deutlicher, daß die Definitionen, die sich in den *Elementen* Euklids finden, ganz und gar nicht geglückt sind.

Es wurde vom 17. Jahrhundert an sehr intensiv diskutiert, ob man für die Grundbegriffe *Wesensdefinitionen* oder *Kausaldefinitionen* (im Sinne von Aristoteles) geben soll, ob man sich mit anderen Arten von Definitionen begnügen kann, ob man die Grundbegriffe überhaupt definieren kann oder ob man sie vielleicht gar nicht definieren muß.

Heron, der vermutlich im ersten Jahrhundert unserer Zeitrechnung in Alexandria lebte, gab in seiner Schrift, die den einfachen Titel *„Definitionen“* trägt, die Definition, daß die Vorstellung einer *Linie* sich aus der Vorstellung *eines im Fluß befindlichen Punktes* ergäbe. Proklos gab in seinem Kommentar zum ersten Buch der *Elemente* Euklids die Definition, daß eine *gerade Linie* durch das *gleichgerichtete und unabgelenkte Fließen eines Punktes* entstehe ([19], S. 292, S. 296). Proklos betonte, daß dabei keine körperliche Bewegung, sondern nur eine vorgestellte Bewegung gemeint sei ([19], S. 296). Aber auch eine solche Definition ist problematisch, weil ungeklärt ist, was ein Punkt ist und insbesondere, wie der Prozeß des (stetigen!) gleichgerichteten Fließens begrifflich zu fassen ist.

Es sieht so aus, als ob exakte und begrifflich einfache Definitionen der geometrischen Grundbegriffe vielleicht gar nicht gegeben werden können. Sind es *„irreduzible“* Begriffe (oder *„Stammbegriffe“* im Sinne von I. Kant)?

3 Das Problem der Definierbarkeit der Grundbegriffe der Geometrie

René Descartes meinte, daß die Grundbegriffe der Arithmetik und der Geometrie nicht durch die Sinne in uns hineingekommen seien, sondern als Ideen von Geburt an in unserer Seele vorhanden seien. Descartes sprach in Anlehnung an Cicero und Thomas von Aquin von *eingeborenen Ideen* (*ideae innatae*). Die Grundbegriffe müssen also nach Descartes' Ansicht nicht definiert werden.

Blaise Pascal hat in einem Essay, der unvollendet blieb und etwa in den Jahren 1655–1658 geschrieben wurde (posthum veröffentlicht) und *De L'Esprit Géométrique* genannt wird, sich ähnlich wie Descartes geäußert.

Pascal meinte, daß man in einer mathematischen Disziplin zuerst diejenigen Begriffe aufsuchen soll, die auch ohne genauere Festlegungen jedem Menschen unmittelbar verständlich sind. Diese Begriffe nannte er *„mots primitifs"*. Beispiele solcher „primitiven Begriffe" sind seiner Meinung nach: „Raum", „Zeit", „Gleichheit", „Zahl", „Existenz", etc. Er hielt sie für transsubjektive Selbstverständlichkeiten, die allen Menschen, die der Sprache mächtig sind, unmittelbar vertraut sind. Die Art dieses Vertrautseins (die über die ratio hinausgeht) nannte Pascal ein *„sentiment du coeur"*, also eine Art natürlichen Prinzipienwissens des Herzens. Das „Herz" spielte bei dem Jansenisten Pascal offenbar die Rolle des platonischen „Nous".

Unter Verwendung der *„mots primitifs"* können alle übrigen Begriffe der Mathematik durch reine Nominal-Definitionen eingeführt werden.

Das große Problem, die Grundbegriffe der Geometrie und der Arithmetik durch geeignete Definitionen einzuführen, umging Pascal mit der etwas verwegenen Bemerkung, diese Grundbegriffe könne man nicht definieren und wären sowieso allen Menschen vertraut.

Die intensiven Diskussionen, die im 16. und 17. Jahrhundert über den Status der Geometrie geführt wurden, hatten aber auch zu Erwägungen geführt, die üblichen euklidischen Definitionen durch Definitionen zu ersetzen, die um *Ursachen* für die Existenz der zu definierenden Objekte erweitert sind, so wie es Aristoteles gefordert hatte. Diese Erwägungen führten allmählich zu einer neuen Auffassung der axiomatischen Methode.

Eine solche neue Auffassung findet man in ersten Ansätzen bei Thomas Hobbes in der ersten Abteilung seiner *Elemente der Philosophie*, die den Titel *De Corpore* trägt (London 1655). Hobbes schlug vor, daß man in der Geometrie diejenigen Objekte, *„die eine Ursache und Erzeugungsweise besitzen"*, eben durch diese Ursachen bzw. Erzeugungsweisen definieren sollte (op.cit., XII, 3). Für die „geraden Linien" und die „Kreise" zitierte er die bekannten heronischen Definitionen (diese Objekte werden durch Bewegungen von Punkten erzeugt, siehe oben). Seinen Vorschlag führte er aber nicht weiter aus.

Ehrenfried Walter von Tschirnhaus hat diesen Vorschlag in seinem Traktat *Medicina mentis, sive artis inveniendi praecepta generalia* (Amsterdam 1687, Leipzig 1695) jedoch systematisch ausgearbeitet und damit dem Begriff der Axiomatik eine neue Gestalt gegeben. Um eine Sache wahrhaft begreifen zu können, muß man sie, seiner Meinung nach, im Geist nachbilden können. Die Definitionen der Grundobjekte sollen daher auch die Methoden zu ihrer Konstruktion (bzw. zu ihrer mentalen Rekonstruktion) enthalten. Definitionen, die das leisten, nannte er *genetisch*. In genetischen Definitionen wird also beschrieben, wie die Objekte, die unter den zu definierenden Begriff fallen, erzeugt (generiert) werden können, und aus der Beschreibung der Erzeugung können alle wesentlichen Eigenschaften der fraglichen Objekte abgelesen werden.

Die Axiomatik besteht bei Tschirnhaus aus einer Liste von Definitionen, nämlich den genetischen Definitionen der Grundobjekte. Aus dieser Liste gewinnt man die sämtlichen Theoreme, indem man zunächst durch Analyse der Definitionen der Grundobjekte auf die Axiome der Theorie stößt und aus diesen dann, wie üblich, mit den Mitteln der Logik die übrigen Sätze ableitet.

Christian Wolff war von der axiomatischen Methode, so wie sie von Tschirnhaus vorgetragen wurde, stark beeinflußt. Er legte sie allen seinen mathematischen Lehr-

büchern zugrunde. Das hat zu einer starken Verbreitung dieser Form der Axiomatik geführt. Wolff hat im ersten Band seiner berühmten *Anfangsgründe aller Mathematischen Wissenschaften* (Leipzig, 1710) einen *„Kurtzen Unterricht von der Mathematischen Methode oder Lehrart“* eingefügt. Er schreibt dort, daß die Prinzipien einer mathematischen Theorie die wohldefinierten Grundbegriffe sind, und daß sich aus den Inhalten dieser Grundbegriffe die Axiome der Theorie (durch Analyse) ergeben und daß aus den Axiomen durch formal-logisches Schließen die Theoreme gewonnen werden können. Dies ist *„die Ordnung, deren sich die Mathematiker in ihrem Vortrage bedienen“* (Wolff, *op. cit.*, § 1).

Die Grundlagen einer mathematischen Disziplin bestehen nach der Tschirnhaus-Wolffschen Auffassung nur aus einer Liste von genetischen Definitionen. Aus den Inhalten dieser Definitionen gewinnt man die Axiome der Disziplin. Die Axiome sind also *ex terminis* gewiß, denn was in ihnen zum Ausdruck kommt, muß sich unmittelbar aus den in den Definitionen niedergelegten Inhalten ergeben.

Betonen möchte ich dabei, daß die Gewißheit der geometrischen Axiome sich hier nicht aus der Anschauung ergibt, sondern aus den Inhalten, die in den genetischen Definitionen niedergelegt sind!

Aber auch die Tschirnhaus-Wolffsche Auffassung hat ihre Schwächen, denn sie bezieht sich auf Erzeugungsprozesse, die oft genug nur schwer (oder gar nicht) begrifflich zu fassen sind. Wie erzeugt man Punkte und wie beschreibt man die Erzeugung einer geraden Linie, einer Fläche etc. mit einfachen Worten, die nicht ihrerseits komplizierte Definitionen erfordern? Es zeigt sich, daß auch die Tschirnhaus-Wolffsche Auffassung im Falle der Geometrie nicht das Gewünschte zu leisten vermag.

Die auf Descartes zurückgehende *Analytische Geometrie*, in der alle geometrischen Begriffe auf algebraische Weise eingeführt werden, konnte schon seit langer Zeit einwandfrei aufgebaut werden. Aber der Wunsch, auch die euklidische (synthetische) Geometrie mit der gleichen Strenge zu entwickeln, war immer noch unerfüllt geblieben. Die vielen Fortschritte auf dem Gebiet der synthetischen Geometrie [die Entstehung der *Projektiven Geometrie* (Victor Poncelet, Jacob Steiner), der *Affinen Geometrie* (August Ferdinand Möbius), die Entdeckung der *Nichteuklidischen Geometrien* (Carl Friedrich Gauss 1816, János Bólyai 1832, Nikolai Lobatschewski 1835), etc.] machten eine sorgfältige Grundlegung immer dringlicher.

Das Ergebnis ist ernüchternd: eine befriedigende Grundlegung der euklidischen Geometrie war auch im 19. Jahrhundert noch nicht gefunden.

4 Die Abnabelung der Geometrie von der Wirklichkeit

Aber am Ende des Jahrhunderts, in den Jahren 1898/1899, war sie dann doch gefunden! Damit dies möglich wurde, mußte der *formale Standpunkt* (auch in der Geometrie) eingenommen werden, um Anschauliches und Inhaltliches abstreifen zu können und eine rein deduktive Geometrie (ohne Berufung auf die Anschauung und Empirie) errichten zu können. Nach vielen Vorarbeiten von Moritz Pasch, Hermann Wiener et al. und insbesondere von Richard Dedekind gelang es David Hilbert 1898/1899, vom formalen Standpunkt aus eine befriedigende Grundlegung der Geometrie zu geben.

Sein Grundgedanke: Geometrie sollte nicht als Naturwissenschaft, die von dem uns umgebenden Raum handelt, entwickelt werden, sondern als Disziplin der Reinen Mathematik. Die Objekte der Geometrie müssen kein raum-zeitliches Substrat besitzen. Hilbert beginnt sein Buch über die *Grundlagen der Geometrie* mit den geflügelten Worten:

„*Wir denken drei verschiedene Systeme von Dingen: die Dinge des ersten Systems nennen wir Punkte und bezeichnen sie mit* $A, B, C, \ldots$; *die Dinge des zweiten Systems nennen wir Gerade und bezeichnen sie mit* $a, b, c, \ldots$; *die Dinge des dritten Systems nennen wir Ebenen und bezeichnen sie mit* $\alpha, \beta, \gamma, \ldots$. [...] *Wir denken die Punkte, Geraden, Ebenen in gewissen gegenseitigen Beziehungen und bezeichnen die Beziehungen durch Worte wie „liegen“, „zwischen“, „parallel“, „congruent“, „stetig“; die genaue und für mathematische Zwecke vollständige Beschreibung dieser Beziehungen erfolgt durch die Axiome der Geometrie.*“

Damit war, wie Hans Freudenthal sich ausdrückte ([7], S. 111, und [8] S. 14),

„*die Nabelschnur zwischen Realität und Geometrie durchgeschnitten. Die Geometrie ist reine Mathematik geworden und die Frage, ob und wie sie auf die Wirklichkeit angewandt werden kann, beantwortet sich bei ihr ganz wie bei irgendeinem anderen Zweige der Mathematik.*“

Es war möglich geworden, neben der euklidischen und den nicht-euklidischen Geometrien auch nicht-archimedische Geometrien, Geometrien, in denen der Satz von Desargues aber nicht der Satz von Pascal gilt, etc. gleichberechtigt aufzubauen und mathematisch zu untersuchen. In der Geometrie Hilberts geht es nicht mehr um außermathematische Inhalte, sondern um die Gestaltung mathematisch-geometrischer Inhalte innerhalb frei gewählter formaler Axiomensysteme.

Mit ähnlichen Worten hat sich auch der Komponist Ernst Krenek (1900–1991) in seinen Vorlesungen *Über Neue Musik* (Wien, 1937) über die hilbertsche Axiomatik der Geometrie geäußert. In seiner 5. Vorlesung „*Musik und Mathematik*“ paraphrasiert er Hilbert wie folgt:

„*Wir denken drei Systeme von Dingen, die Dinge des ersten nennen wir Töne, die Dinge des zweiten nennen wir Akkorde, die Dinge des dritten Melodien. Wir denken die Dinge der drei Systeme in verschiedenen Beziehungen, die wir mit Worten wie Intervall, Konsonanz, Dissonanz, Bewegung, Umkehrung und anderen bezeichnen.*“

Krenek fügt hinzu, daß damit die Autonomie der Musik praktisch ausgesprochen sei ([15], S. 83). Krenek will damit sagen, daß es in der Musik nicht um die Nachbildung außermusikalischer Inhalte geht, sondern um die Gestaltung musikalischer Inhalte innerhalb frei gewählter formaler Systeme. Ein solches System kann neben den historisch überlieferten Systemen auch das System der seriellen 12-Ton Musik sein.

Das hilbertsche Diktum „*Wir denken uns verschiedene Systeme von Dingen ...*“ hat viele Mathematiker beeindruckt und ist oft kopiert worden. Hilbert selbst hat es auch in seinem Essay „*Über den Zahlbegriff*“ [12] verwendet. Es heißt dort:

„Wir denken ein System von Dingen: wir nennen diese Dinge Zahlen, etc.“

Hilbert hat sich mit dieser Sprechweise an Hermann Wiener ([25]) angelehnt und sie von 1894 an immer wieder verwendet (cf. Hallett-Majer, [10], S. 72, 74, 224, 304, 437). Es soll damit zum Ausdruck kommen, daß es auf die spezielle Natur der betrachteten Gegenstände nicht ankommt, und daß es für die Mathematik nur relevant ist, daß die Gegenstände irgendwelche *„Dinge unterschiedlicher Arten“* [von einerlei, zweierlei oder dreierlei Art…] sind, die gewissen Postulaten genügen.

Hilbert hat sich mit dieser Sprechweise aber insbesondere an Richard Dedekind angelehnt, der im Vorwort seines Essays *Was sind und was sollen die Zahlen* (1888, Nachdruck in [2], S. 335–391) die natürlichen Zahlen als

„freie Schöpfungen des menschlichen Geistes“

nachweisen wollte. Genauso möchte auch Hilbert die Gegenstände der Geometrie als freie Schöpfungen des menschlichen Geistes einführen. Die Gegenstände der Geometrie sind keine Gegenstände, die in der natürlichen Umwelt vorgefunden werden können. Sie werden durch Abstraktion oder Idealisierung gewonnen, wobei das Maß der Abstraktion oder Idealisierung von den Mathematikern frei gewählt werden kann.

Wir wenden uns jetzt – wie in der Einleitung angekündigt – den logischen Problemen zu, die die Grundlegung der Geometrie aufwirft.

5 Implizite Definitionen in Hilberts *Grundlagen der Geometrie*

Im Unterschied zu Euklid, Aristoteles, Tschirnhaus, Wolff et al. stehen bei Hilbert zu Anfang seines Axiomen-Systems keine (expliziten) Definitionen der Grundbegriffe. Hilbert sagt nur, daß er über drei verschiedene Sorten von Gegenständen sprechen will und daß er dabei zwei 2-stellige Relationen und eine 3-stellige Relation verwenden will. Die *intendierte Interpretation* ist, daß es sich dabei um

Punkte, gerade Linien (unbegrenzter Länge) und *Flächen*,
sowie die *Inzidenz-*, die *Kongruenz-* und die *Zwischen-Relation*

handeln soll. Eine inhaltliche Beschreibung (Definition) dieser Objektbereiche und Relationen wird aber nicht vorab gegeben. Es wird also nicht vorab gesagt, was *„Punkte“* sind, was *„Linien“* und was *„Flächen“* sind, und es wird auch nicht vorab gesagt, was *„Inzidenz“*, *„Kongruenz“* und *„Zwischen“* bedeutet. In den Axiomen, in denen diese Objektbereiche und Relationen angesprochen werden, wird nur festgelegt, in welchen Beziehungen sie untereinander stehen und wie mit ihnen umgegangen werden kann. Die Möglichkeit, diese Objektbereiche und Relationen zu interpretieren, wird aber damit eingeengt. Diese Einengung (Eingrenzung) ist eine Form von „Definition“, die bereits von Joseph-Diaz Gergonne 1818 als *„implizite Definition“* bezeichnet wurde.

In den Axiomen wird nur das ausgesprochen, was überhaupt begrifflich faßbar ist und mathematisch relevant ist, und alles übrige, was sonst noch seit alters her zum Wesen und Erscheinungsbild der geometrischen Objekte und Relationen gehört,

weggelassen (z.B., daß Kreise wirklich *rund* sind, daß Geraden wirklich *gerade* sind, daß kongruente Winkel *gleich groß* sind, etc.).

Man darf also sagen, daß in der hilbertschen Axiomatik der Geometrie die Grundobjekte und Grundbegriffe *nicht explizit* definiert werden, und daß damit eine Ähnlichkeit zur descartesschen und pascalschen Auffassung vorliegt. Aber im Unterschied zu Descartes und Pascal werden *implizite* Definitionen gegeben, damit klar ist, was mathematisch relevant ist. Inhaltliche Bestimmungen, die über das mathematisch Relevante hinausgehen, werden nicht gegeben.

Die Axiome der Geometrie haben bei Hilbert das folgende Aussehen (bei Hilbert allerdings ohne Verwendung logischer Zeichen), beispielsweise:

$$\text{Axiom (I.1):} \quad \forall A\, \forall B\, \exists a : I(A,a) \& I(B,a),$$

wobei I ein (leeres) Zeichen für eine 2-stellige Relation ist, deren intendierte Interpretation die Inzidenzrelation ist. Das Axiom hat also die intendierte Interpretation: Zu je zwei Punkten A und B gibt es eine Gerade a, auf der die beiden Punkte A und B liegen.

Neben diesem Axiom (I.1) gibt es im Falle der zwei-dimensionalen Geometrie noch zwei weitere *„Axiome der Verknüpfung“* (I.2) (I.3), vier *„Axiome der Anordnung“* (II.1), ..., (II.4), in denen der Begriff „ein Punkt liegt *zwischen* zwei anderen Punkten“ implizit definiert wird, fünf *„Kongruenz-Axiome“* (III.1), ..., (III.5), in denen die Kongruenz von Strecken und die Kongruenz von Winkeln implizit definiert wird, das euklidische *Parallelen-Axiom* (IV), und schließlich das *„archimedische Axiom“* (V.1) und das *„Axiom der Vollständigkeit“* (V.2).

Gottlob Frege war mit dem hilbertschen Zugang zur Geometrie gar nicht einverstanden. Frege vertrat den *inhaltlichen Standpunkt* und wollte implizite Definitionen nicht gelten lassen. Er meinte (genauso wie Descartes und Pascal), daß die Grundbegriffe der Geometrie zwar undefinierbar seien, daß aber ihr Inhalt dennoch allen Menschen vertraut sei und daß man sich in der Geometrie auf diesen Inhalt berufen könne (vergl. Freges Essay *Über die Grundlagen der Geometrie*, Teil I, (Jahresberichte der DMV 15 (1906), S. 293–309; Nachdruck in [6], S. 292).

Hilbert verteidigte seinen *formalen Standpunkt* und schrieb Frege:

> *„Jedes Axiom trägt ja zur Definition etwas bei und jedes neue Axiom ändert also den Begriff. ‚Punkt‘ in der Euklidischen, Nicht-Euklidischen, Archimedischen, Nicht-Archimedischen Geometrie ist jedesmal etwas Anderes.“*

(Hilbert in einem Brief an Frege vom 29. Dez. 1899).

Über den anschaulichen Gehalt des Begriffes *‚Punkt‘* in den verschiedenen Geometrien will Hilbert keine Aussage machen. Hilbert behauptet, daß der Begriff des *‚Punktes‘* in den verschiedenen Geometrien verschiedene Merkmale hat. (In diesen Merkmalen wird ja auch auf Linien etc. Bezug genommen!) Auf einen eindeutig bestimmten „Inhalt“ der Begriffe *‚Punkt, Gerade, Winkel‘* etc. kann man sich also in den verschiedenen Geometrien nicht berufen. Mit dieser Einsicht unterscheidet sich Hilbert von allen seinen Vorgängern, insbesondere auch von Pasch.

Albert Einstein hat sich nachdrücklich zu dieser formalen Auffassung der Geometrie bei Hilbert bekannt. In seinem Vortrag über *Geometrie und Erfahrung* ([3], S. 124–125) heißt es über die hilbertsche Grundlegung der Geometrie:

„Die Geometrie handelt [hier] *von Gegenständen, die mit den Worten Gerade, Punkt usw. bezeichnet werden. Irgendeine Kenntnis oder Anschauung wird von diesen Gegenständen nicht vorausgesetzt, sondern nur die Gültigkeit jener ebenfalls rein formal, d.h. losgelöst von jedem Anschauungs- und Erlebnisinhalte, aufzufassenden Axiome… Diese Axiome sind freie Schöpfungen des menschlichen Geistes. Alle anderen geometrischen Sätze sind logische Folgerungen aus den (nur nominalistisch aufzufassenden) Axiomen. Die Axiome definieren erst die Gegenstände, von denen die Geometrie handelt… Diese von der modernen Axiomatik vertretene Auffassung der Axiome säubert die Mathematik von allen nicht zu ihr gehörenden Elementen und beseitigt so das mystische Dunkel, welches der Grundlage der Mathematik bisher anhaftete… Dieser geschilderten Auffassung der Geometrie lege ich deshalb besondere Bedeutung bei, weil es mir ohne sie unmöglich gewesen wäre, die Relativitätstheorie aufzustellen.“*

Die hilbertsche Begründung der Geometrie vom formalen Standpunkt aus war für Einstein insofern von erheblicher Bedeutung, weil in ihr der anschauliche Gehalt vom logisch-formalen Gehalt sauber getrennt war und so auch neue Interpretationen möglich wurden.

Die hilbertsche Geometrie und die „Moderne Algebra“ [von Heinrich Weber (Math. Ann. 43 (1893), S. 521–549), Emil Artin, Emmy Noether, Bartel van der Waerden et al.] haben sich ganz wesentlich auf den Begriff der ‘Axiomatik’ gestützt. Aber unter ‘Axiomatik’ verstand man jetzt etwas anderes als von Aristoteles an bis ins ausgehende 19. Jahrhundert. In den Axiomen tauchen jetzt Zeichen für Begriffe auf, für die keine expliziten Definitionen gegeben werden, die daher auch keine *Symbole* sondern nur *leere Zeichen* sind,[7] und die lediglich durch die Gesamtheit der Axiome *implizit definiert* werden.

6 Das hilbertsche Vollständigkeits-Axiom

Im hilbertschen Axiomensystem findet sich ein „Axiom“, das eine ganz andere Funktion hat, als alle übrigen Axiome. Es ist das sogenannte „Vollständigkeits-Axiom“:

(V.2) Axiom der Vollständigkeit: „*Die Elemente* (*Punkte, Geraden, Ebenen*) *der Geometrie bilden ein System von Dingen, welches bei Aufrechterhaltung sämtlicher genannten Axiome keiner Erweiterung mehr fähig ist*.“

Dieses Axiom hat die Modelle der übrigen Axiome zum Gegenstand. Es sagt, daß das System der Punkte, Geraden und Ebenen, das *„wir denken“* sollen, ein *maximales Modell* der übrigen Axiome ist. Hilbert zeigt, daß die maximalen Modelle untereinander alle isomorph sind und daß sie insbesondere mit der 3-dimensionalen

[7] Das griechische Wort ‘Symbolon’ ist aus den Wörtern ‘syn’ (zusammen) und ‘ballein’ (werfen) zusammengesetzt. Ein *Symbol* ist demnach ein *Zusammengefügtes*, d.h. ein Zeichen, dem eine bestimmte Bedeutung beigelegt ist. Ein Gegenstand oder eine Figur wird zu einem Symbol, wenn ihm ein bestimmter Sinn beigelegt wird. Man muß in die Symbolik eingeweiht sein, um den beigelegten Sinn erkennen zu können.

Ein Zeichen, dem keine Bedeutung beigelegt ist, ist ein *leeres Zeichen*.

cartesischen analytischen Geometrie (über dem Körper $\mathbb{R}$ der reellen Zahlen) isomorph sind. Das Axiomen-System ist also kategorisch (d.h. monomorph). Damit ist das große Ziel erreicht und eine axiomatische Kennzeichnung der euklidischen Geometrie gefunden.

Ähnlich wie Dedekind 1888 zeigen konnte, was aus mathematischer Sicht „die natürlichen Zahlen“ sind, so wollte Hilbert zeigen, was aus mathematischer Sicht „die euklidische Geometrie“ ist. Dedekind konnte eine Familie von Strukturen angeben und zeigen, daß in dieser Familie die *minimalen* Strukturen alle paarweise isomorph sind und daß dieser Isomorphie-Typ aus mathematischer Sicht den Bereich der natürlichen Zahlen charakterisiert (cf. Felgner [5], S. 39). Ganz analog konnte Hilbert zeigen, daß in der Familie aller Modelle seiner Axiome (I.1), ..., (V.1) die *maximalen* Strukturen (in ihnen gilt auch (V.2)) alle paarweise isomorph sind und daß dieser Isomorphie-Typ die euklidische Geometrie charakterisiert.

Es ist bemerkenswert, daß das „Axiom der Vollständigkeit“ (V.2) in der Festschrift, die im Juni 1899 publiziert wurde, noch nicht auftritt. Aber Hilbert muß es noch im selben Monat aufgestellt haben, wie aus einem Brief von Hermann Minkowski vom 24. 6. 1899 hervorgeht (cf. Hallett-Majer [10], S. 433, oder Toepell [24], S. 254). Auch in der französischen Übersetzung der Festschrift (publiziert 1900 in Band 17 der Annales Sci. Ecole Normale Sup., S. 103–209) kommt es bereits vor. Wie Edmund Husserl überliefert hat, hat Hilbert über das „Axiom der Vollständigkeit“ am 5. November 1901 in der Göttinger Mathematischen Gesellschaft vorgetragen (cf. Husserl [13], S. 444–451). In der zweiten Auflage der *Grundlagen der Geometrie* (Leipzig 1903, S. 16) tritt es dann auch als Axiom (V.2) auf.

Hilbert hat in dem genannten Vortrag am 5. XI. 1901 darauf hingewiesen, daß das archimedische Axiom (V.1) benötigt wird, um sicherzustellen, daß es auf jeder Geraden ein dichtes Netz rationaler Zahlen gibt. Das „Vollständigkeits-Axiom“ (V.2) liefert dann, daß jede Gerade die Struktur der reellen Zahlen trägt. In Gegenwart des archimedischen Axioms ist also das „Vollständigkeits-Axiom“ widerspruchsfrei mit allen übrigen Axiomen und garantiert, daß die Geometrie mit der klassischen Geometrie in drei Dimensionen über dem Körper $\mathbb{R}$ der reellen Zahlen isomorph ist. In Abwesenheit des archimedischen Axioms ist allerdings das „Vollständigkeits-Axiom“ mit den restlichen Axiomen unverträglich, denn jeden formal-reellen Körper kann man (wie Emil Artin und Otto Schreier, Hamburger Abhandlungen 5 (1926), S. 85–99, gezeigt haben) durch Adjunktion eines transzendenten Elementes zu einem größeren formalreellen Körper (der aber nicht mehr archimedisch geordnet ist) erweitern.

Max Dehn hat in seiner Besprechung der 2. Auflage (1903) der hilbertschen *Grundlagen der Geometrie* im Jahrbuch über die Fortschritte der Math., Band 34 (Jahrgang 1903, Berlin 1905, S. 523) bemerkt, daß das Vollständigkeits-Axiom (V.2) *„in seinen Anwendungen etwa mit dem dedekindschen Axiom von der Existenz der Grenze äquivalent“* ist.

Federigo Enriques hat in seinem Bericht über die *Prinzipien der Geometrie* in der Encyclopädie der Mathematischen Wissenschaften (3. Band, Heft 1, 1907) die Aussage von Dehn bestätigt und überdies darauf hingewiesen, daß nach Otto Stolz (Math. Annalen 22 (1883), S. 504–519) und Otto Hölder (Ber. Verh. Sächs. Ges. Wiss. Leipzig, 53 (1901), S. 1–64) auf der Grundlage der übrigen Axiome aus dem

dedekindschen Axiom das archimedische Postulat folgt. Das ***dedekindsche Axiom*** kann man etwa wie folgt formulieren:

> *Zu beliebigen Mengen X und Y von Punkten, derart, daß alle Punkte aus X zwischen einem Punkt a und allen Punkten aus Y liegen, gibt es immer einen Punkt b, der zwischen allen Punkten aus X und allen Punkten aus Y liegt.*

$$\forall X\ \forall Y\ \Big[\big(\exists a\ \forall x \in X\ \forall y \in Y : \zeta(a,x,y)\big)\ \Rightarrow\ \exists b\ \forall x \in X\ \forall y \in Y : \zeta(x,b,y)\Big],$$

wenn ζ die „Zwischen-Relation“ bezeichnet: $\zeta(x,y,z)$, mit der intendierten Interpretation: „y liegt zwischen x und z“. Nach Dehn, Hölder und Stolz sind also die beiden hilbertschen Stetigkeitsaxiome, nämlich das archimedische Axiom (V.1) und das Vollständigkeitsaxiom (V.2), zusammen genommen auf der Basis der übrigen Axiome mit dem obigen dedekindschen Stetigkeitsaxiom äquivalent. Diese Ersetzung stellt eine erhebliche Vereinfachung des hilbertschen Axiomensystems dar, denn das archimedische Axiom, das in seiner Formulierung den Begriff der natürlichen Zahl verwendet, und das hilbertsche Vollständigkeitsaxiom, das sich auf den Begriff des Modells eines Axiomensystems stützt, werden ersetzt durch ein einziges Axiom, das sich nur auf geometrische Grundbegriffe und den Mengenbegriff stützt. Da es sich auf den Begriff der Menge (von Punkten) stützt, ist es kein elementares Axiom, sondern ein Axiom, das in einer Sprache der 2. Stufe formuliert ist.

Das Vollständigkeitsaxiom (V.2) ist, wie Arnold Schmidt sich ausgedrückt hat, ein *„Axiom über Axiome“* ([20], S. 407), oder wie Ivor Grattan-Guinness schrieb *„a kind of meta-axiom“* ([9], S. 210). Freudenthal hielt es für ein *„unglückseliges Axiom“* ([7], S. 117). Alle anderen Axiome sprechen darüber, was *in* den einzelnen geometrischen Strukturen gelten soll, aber das Vollständigkeitsaxiom spricht *über* die Klasse aller geometrischen Strukturen; es soll aus der Klasse aller geometrischen Strukturen die maximalen Strukturen herausheben. Insofern fällt das Vollständigkeitsaxiom aus dem Rahmen und ist kein eigentliches Axiom.[8] Es sollte (zusammen mit dem archimedischen Axiom) durch das dedekindsche Axiom ersetzt werden. Das so modifizierte Axiomensystem ist nach wie vor kategorisch und vollständig,[9] aber überdies ausgesprochen leicht zu handhaben.[10] Alle Propositionen von Euklids *Elementen* lassen sich hier sehr elegant und einwandfrei beweisen.

---

[8] Auch in anderen mathematischen Disziplinen wurden derartige meta-mathematische Vollständigkeits-Axiome aufgestellt, etwa von Adolf Abraham Fraenkel (vergl. seine *Einleitung in die Mengenlehre*, Berlin 1923, S. 219), Paul Finsler (Math. Z. 25 (1926), S. 683–713) et al. Sie wurden von Reinhold Baer (Math. Z. 27 (1928), S. 536–539, S. 543), Richard Baldus (Math. Annalen 100 (1928), S. 321–333), Paul Bernays (Math. Z. 63 (1955), S. 219–229), Rudolf Carnap und Friedrich Bachmann (Erkenntnis 1 (1930/31), S. 303–307, Erkenntnis 6 (1936), S. 166–188) et al. ausführlich diskutiert. Die Kritiker wenden ein, daß die gewöhnlichen Axiome es erlauben, Beweise unter Anwendung rein deduktiver Logik-Kalküle zu führen, daß aber die meta-mathematischen Vollständigkeits-Axiome nur ein semantisches Vorgehen (Nachweis der Erfüllbarkeit in den maximalen Strukturen) gestatten.

[9] Hilbert selbst hat auch in allen späteren Auflagen seiner *Grundlagen der Geometrie* an seiner ursprünglichen Axiomatisierung festgehalten. Andere Autoren (beispielsweise Alfred Tarski, [22, 23]) haben es jedoch vorgezogen, die beiden hilbertschen Stetigkeitsaxiome (V.1) und (V.2) durch das dedekindsche Axiom zu ersetzen.

[10] In dem hilbertschen Axiomensystem kann man die Systeme der „Geraden“ und der „Ebenen“ weglassen, denn jede Gerade ist durch zwei verschiedene Punkte, die auf ihr liegen, eindeutig bestimmt und über alle weiteren auf ihr liegenden Punkte kann unter Verwendung der „Zwischen“-Relation gesprochen werden.

Abschließend wollen wir noch – wie in der Einleitung angekündigt – auf die ontologischen Probleme, die die Geometrie aufwirft, eingehen.

7 Hilberts Strukturalismus

Wir hatten schon in der Einleitung die kritische Frage gestellt, was denn „Punkte", „gerade Linien" etc. ihrem Wesen nach sind. Da sie ausdehnungslos, bzw. breitenlos sein sollen, können sie nicht sinnlich wahrnehmbar sein und können auch nicht wirklich vorgestellt werden. Man darf mit Recht fragen, ob es sie wirklich gibt, wo es sie gibt, oder in welchem Sinne es sie gibt.

Über die euklidische Definition des Begriffes *„Punkt"* schrieb der italienische Mathematiker Nicolo Tartaglia in seiner Übersetzung der euklidischen *Elemente* (Venedig 1543), daß man Punkte nirgendwo in der Welt finden kann und sie sich auch gar nicht vorstellen kann. Er schrieb:

„*... non si puo tuoglier, ne trouar, ne anchora imaginar la mettade.*"

Der englische Philosoph Thomas Hobbes hielt Linien ohne Breite für etwas unbegreifliches:

„Lineam sine latitudine, rem inconceptibilem."

Ähnliche Aussagen findet man in der kritischen Literatur immer wieder, aber kein Mathematiker konnte von der Antike an bis ins ausgehende 19. Jahrhundert etwas Überzeugendes dagegen setzen. Die englisch-irisch-schottischen Empiristen im 18. und 19. Jahrhundert glaubten das Problem lösen zu können.

Für David Hume beispielsweise sind „mathematische Punkte" solche Raumgrößen, die dem Gesichtssinn und dem Tastsinn als unteilbar erscheinen. Er meint, daß es solche Raumgrößen gibt. Zum „Beweis" sagt er, daß Tintenklekse auf einem Stück Papier, die man aus hinreichend großer Entfernung anschaut, wo sie eben noch sichtbar sind, dem Auge als unteilbare Gegenstände erscheinen (D. Hume: *Philosophical Essays concerning human understanding*, London 1748).

John Stuart Mill definierte in seinem Buch *A System of Logic* (London 1843, S. 148) den mathematischen Punkt ebenfalls als

„the minimum visible, the smallest portion of surface we can see,"

und auch noch Moritz Pasch schrieb in seinen berühmten *Vorlesungen über neuere Geometrie* (Leipzig, 1882):

„Allemal aber werden die Körper, deren Teilung sich mit den Beobachtungsgrenzen nicht verträgt, „Punkte" genannt" ([16], S. 3).

Für Ebenen gilt eine analoge Aussage. Darauf haben B. Levi (Torino Mem, 1904, S. 283) und O. Veblen (AMS Trans. 5 (1904), S. 343) hingewiesen. W. Schwabhäuser, W. Szmielew & A. Tarski (cf. [21]) haben diese Vereinfachung ihren modelltheoretischen Untersuchungen der Geometrie zugrunde gelegt. In der Tat sind einsortige Kalküle vom Standpunkt der Logik aus einfacher zu handhaben als mehrsortige Kalküle.

Es sollte eigentlich allen Mathematikern und Philosophen klar sein, daß die Geometrie, wenn sie eine mathematische Disziplin sein soll, nicht von solchen unscharf definierten Dingen handeln kann. Andernfalls wäre die Geometrie lediglich eine Wissenschaft, in der es nur um ungefähr gültige Aussagen geht.

Wir entnehmen dieser Beispielsammlung, daß es ja gar nicht so klar zu sein scheint, ob es überhaupt „mathematische Punkte“, „breitenlose Linien“ etc. wirklich gibt. Es stellt sich die Frage, worüber die Geometer überhaupt reden.

Auch das analoge Problem, worüber man überhaupt in der Arithmetik redet, blieb ebenfalls lange Zeit unklar. Es war schon in der Antike gestellt worden und seither von Mathematikern und Philosophen kontrovers diskutiert worden. Was sind eigentlich die natürlichen Zahlen? Erst Dedekind war 1888 eine überzeugende Lösung gelungen.

Nach Dedekind kann die Frage, welchen ontologischen Status die natürlichen Zahlen haben, nur im Sinne des *Strukturalismus* (den er begründete) beantwortet werden. Es gibt keine Dinge, die *per se* als natürliche Zahlen angesehen werden können. Die natürlichen Zahlen lassen sich nur als Elemente von unendlichen Bereichen erklären, die gewissen Rechenregeln genügen. Nach Dedekind kann ein Ding beispielsweise den Namen „17“ tragen, wenn es einem Rechenbereich angehört, der bestimmten Gesetzen genügt, etwa den Dedekind-Peanoschen Axiomen, und wenn es in diesem Bereich die entsprechende Stellung hat.

Hilbert war vom Dedekindschen Zugang zum Zahlbegriff tief beeindruckt. Er erkannte in den Jahren 1898–1900, daß auch in der Arithmetik der reellen Zahlen und in der Geometrie ähnlich lautende Lösungen möglich sind. Im Falle der Geometrie lautete seine Lösung wie folgt:

Ein Objekt kann als (ausdehnungsloser) „Punkt“ oder als (breitenlose) „Linie“ angesehen werden, wenn es einem *Gefüge* von Objekten angehört, die alle zusammen die euklidischen (oder hilbertschen) Axiome der Geometrie erfüllen. Diese Objekte müssen nicht im umgangssprachlichen (inhaltlichen) Sinne „ausdehnungslos“, bzw. „breitenlos“ sein; sie müssen es nur im Sinne des Gefüges sein.

„Gefüge“ heißt im Lateinischen structura. Eine „Struktur“ ist im wörtlichen Sinne eine Menge von „ordentlich zusammengefügten“ Dingen.[11] In der Mathematik ist es seit etwa 1900 üblich geworden, eine Menge als „strukturierte Menge“ zu bezeichnen, wenn in ihr einige Relationen und einige Operationen ausgezeichnet sind. „Strukturierte Mengen“ nennt man auch kurz „Strukturen“. (Dedekind, Hilbert, Weber et al. sprachen statt dessen von *„Systemen“*, ohne jedoch diesen Begriff als terminus technicus einzuführen.)

In den oben angegebenen Beispielen wurde deutlich, daß die Eigenschaften

- beispielsweise die „siebzehnte natürliche Zahl“ zu sein,
- oder eine „breitenlose Linie“ zu sein, etc.

jeweils Eigenschaften sind, die Objekte nicht für sich ganz allein (*per se*) besitzen, sondern nur *als* Elemente geeigneter Strukturen.

Die Auffassung, daß eine nicht-triviale mathematische Eigenschaft einem Objekt nur dann zukommt, wenn es in eine passende Struktur eingebettet ist und dort die

[11] „Structura“ ist im Lateinischen die „Schichtung, die ordentliche Zusammenfügung, das Mauerwerk“.

fragliche Eigenschaft hat, nennt man *Strukturalismus*. Man kann sich auf mathematische Objekte nur dann beziehen, wenn man auch die Strukturen, denen diese Objekte angehören sollen, angibt.

Die Frage, was die verschiedenen geometrischen Objekte „sind", wo es sie gibt und in welchem Sinne es sie gibt, ist im Strukturalismus *erledigt*, denn all diese Objekte haben die jeweiligen Eigenschaften immer nur innerhalb geeigneter Strukturen und es gibt all diese Objekte auch immer nur innerhalb von Strukturen. – Zur Konstruktion mathematischer Strukturen ist allerdings etwas Mengenlehre nötig.

Die Kennzeichnung des Bereichs aller natürlichen Zahlen als minimale Struktur, in der die bekannten Dedekind-Peanoschen Axiome gelten (durch Dedekind 1888), und die Kennzeichnung der euklidischen Ebene und des euklidischen Raumes als maximale Strukturen, in denen die hilbertschen Axiome gelten (durch Hilbert, 1899), sind berühmte Beispiele für die strukturalistische Auffassung in der Mathematik.

8 Geometrie und Wirklichkeit

Hilbert hat seinem Buch ein treffendes Motto vorangestellt:

> *„So fängt denn alle menschliche Erkenntnis mit Anschauungen an, geht von da zu Begriffen und endigt mit Ideen."*

Das Motto ist Immanuel Kants *Kritik der reinen Vernunft* (Elementarlehre, II. Teil, II. Abteilung, A 702, B 730) entnommen. An einer anderen Stelle (A 298, B 355) spricht Kant diesen Gedanken mit ähnlichen Worten aus:

> *„Alle unsere Erkenntnis hebt von den Sinnen an, geht von da zum Verstande und endigt bei der Vernunft."*

Dieses Motto ist eine gute Kennzeichnung der Zielsetzung, die Hilbert in seinem Buch über die *Grundlagen der Geometrie* verfolgte: die ursprünglich nur anschaulich gegebenen Objekte der Geometrie sollen in idealisierter Form begrifflich gefaßt werden. Die Geometrie soll sich nicht mehr auf die Anschauung berufen müssen, damit sie als reines *„Fachwerk von Begriffen"*[12] der mathematischen Vernunft vorgelegt werden kann.

Der Geometrie liegt ein reiches Material empirisch gefundener Einsichten zugrunde, aber typisch für das Vorgehen in der Mathematik ist, daß in einem zweiten Schritt das anschaulich Gegebene in Begriffe umgesetzt wird, um so das anschauliche Erkennen in ein logisches, schrittweises Folgern zerlegen zu können. – Wir haben darüber bereits in Abschn. 1 gesprochen und kommen jetzt am Ende unseres Essays wieder auf diese grundsätzlichen Probleme zurück.

Damit aber die logische Analyse vollständig und einwandfrei ist, muß die Geometrie von allem *„Erdenrest"* befreit werden. Das Wort *„Erdenrest"* paßt hier nicht schlecht. Es ist von Johann Wolfgang Goethe im *Faust II* (5. Akt, Zeile 11954) in

[12] So hat sich Hilbert in seinen Vorlesungen (vergl. Hallett-Majer [10], S. 72, 540), in seinem Brief vom 29. Dezember 1899 an Frege, in seinem Aufsatz *Axiomatisches Denken* (1918) und in seiner Vorlesung *Natur und mathematisches Erkennen* (1919, S. 43) ausgedrückt.

die deutsche Sprache eingeführt worden und sollte eine nur teilweise vollzogene Abstraktion bezeichnen. Die vollendeteren Engel, die keine reinen Geistwesen sind und denen noch manches allzu Menschliche anhaftet, singen dort:

„Uns bleibt ein Erdenrest // Zu tragen peinlich,“... etc.

Nach Goethe haben auch Friedrich Nietzsche (in *Jenseits von Gut und Böse*, 1. Hauptstück, Nr. 17), Felix Hausdorff (in seinen *Grundzügen der Mengenlehre*, Leipzig, 1914, S. 45) und vielleicht auch ein paar andere Schriftsteller dieses schöne und seltene Wort benutzt. Auch Otto Blumenthal hat es in seiner *Hommage an David Hilbert* zu seinem 60. Geburtstag (erschienen in: „Die Naturwissenschaften“, Band 10 (1922), S. 67–72) verwendet. Er schrieb dort über

„... die von Kronecker vertretene und von Hilbert immer leidenschaftlich bekämpfte Auffassung, daß aller Mathematik, die sich nicht unmittelbar an die ganze Zahl anknüpfen lasse, ein unreinlicher Erdenrest anhafte.“

Nach Leopold Kronecker (Werke III,1, S. 253, 274) müssen alle Ergebnisse der reinen Mathematik letztlich *„in einfachen Formen der Eigenschaften ganzer Zahlen ausdrückbar sein“* und sollen sich nirgendwo auf sinnliche Anschauung berufen müssen. Die Gefahr, daß in geometrische Argumentationen sinnliche Anschauung einfließt, ist sehr groß und nach Kronecker nur dann behoben, wenn die Geometrie arithmetisiert wird (wie es in der cartesischen *analytischen* Geometrie ja auch der Fall ist). Hilbert war nach dem Zeugnis seines Schülers Blumenthal zutiefst davon überzeugt, daß auch die *synthetische* euklidische Geometrie einen Aufbau gestattet, der sich nirgendwo auf die sinnliche Anschauung stützen muß, daß sie also auch ohne Arithmetisierung vollständig von allem *Erdenrest* befreit werden kann. Daß dies möglich ist, sollte die Schrift *Grundlagen der Geometrie* zeigen.

Deshalb begann Hilbert seine Schrift auch mit dem etwas provokanten Satz:

„Wir denken drei verschiedene Systeme von Dingen, ...“

Er verweist nicht auf die Wirklichkeit, wie es noch Moritz Pasch in seinen *Vorlesungen über die neuere Geometrie* (Berlin 1882) getan hat, sondern bezieht sich ganz bewußt auf Objekte, die nur gedacht werden, denen also kein *Erdenrest* mehr anhaftet. Auch die Grundbegriffe werden nicht durch Angabe ihrer anschaulichen Inhalte eingeführt, sondern rein formal durch implizite Definitionen. Auf diese Weise ist die Geometrie von allem *Erdenrest* befreit worden und es ist möglich geworden, die Geometrie einer logischen Analyse zu unterziehen. Damit ist die Geometrie, wie Freudenthal ([8], S. 14) betont hat, reine Mathematik geworden.

Aber ihre Herkunft als Wissenschaft, die vom sinnlich wahrnehmbaren Raum handelt, soll und kann die Geometrie keineswegs leugnen.

- Als *„allgemeines Wissensgebiet“* ist die Geometrie für Hilbert eine *„Naturwissenschaft“* (cf. Hallett-Majer [10], S. 72, 221, 302, 540).
- Aber die *„Theorie des Wissensgebietes“* ist nach Hilbert das *„Fachwerk der Begriffe“* (vergl. Hilbert: *Axiomatisches Denken*, 1918). Die *„Theorie des Wissensgebietes“* ist keine Naturwissenschaft, sie ist ein Gebiet der reinen Mathematik.

Als *allgemeines Wissensgebiet* stützt sich die Geometrie auf die Anschauung und geht *„auf ein lebendiges Erfassen der Gegenstände und ihrer inhaltlichen Beziehungen aus“* (Hilbert & Cohn-Vossen: *Anschauliche Geometrie*, 1932, Vorwort). Aber

dem lebendigen, anschaulichen Erfassen sind Grenzen gesetzt und wie in jeder Naturwissenschaft gibt es auch in einer Geometrie, die sich nur auf die Anschauung stützt, viele unlösbare Probleme.

In der *Theorie des Wissensgebietes* dürfen jedoch *Idealisierungen* vorgenommen werden, die das empirisch Wahrnehmbare übersteigen. In seinen *Grundlagen der Geometrie* hat Hilbert einige Idealisierungen vorgenommen. (Es handelt sich um Idealisierungen und nicht um Abstraktionen!) So sind die geraden Linien von vornherein unbegrenzt. Diese geraden Linien sollen wie in der euklidischen Geometrie keine Breite haben. Auch das zweite Anordnungs-Axiom (II.2) drückt eine Idealisierung aus, indem es die unendliche Teilbarkeit einer jeden vorgegebenen Strecke aussagt. Die beiden Stetigkeitsaxiome (V.1) und (V.2) lassen sich ebensowenig auf empirischem Wege begründen.

Mit der Sprechweise *„wir denken uns verschiedene Systeme von Dingen . . .“* soll zum Ausdruck kommen, daß die einzelnen geometrischen Gegenstände nicht erst, einer nach dem anderen, konstruiert werden müssen, sondern daß sie Systeme (d.h. Mengen) bilden, die gleich von Anfang an vollständig gegeben sind (und insofern aktual unendliche Mengen bilden). In seinem Vortrag *„Über den Zahlbegriff“* ([12], S. 181) sagt es Hilbert in aller Deutlichkeit, daß man

> *„beim Aufbau der Geometrie. . . mit der Annahme der Existenz der sämtlichen Elemente zu beginnen“*

pflegt. Das hat viele Konsequenzen, insbesondere auch die, daß es gute Gründe gibt, *imprädikative Definitionen* und die üblichen Gesetze der klassischen Quantorenlogik der 1. Stufe zuzulassen (beispielsweise das *tertium non datur*: $\neg\forall x : \Phi(x) \Rightarrow \exists x : \neg\Phi(x)$).

Damit läßt sich auch die eigentliche Zielsetzung der axiomatischen Methode bei Hilbert klar und deutlich umreißen: die sämtlichen Idealisierungen, die einer mathematischen Theorie zugrunde gelegt werden, werden im System der Axiome hinterlegt. Aus den Axiomen werden die Theoreme durch rein formal-logisches Schließen gewonnen (also ohne Verwendung weiterer Idealisierungen und ohne Berufung auf die Anschauung). Diesen Charakter des mathematischen Schließens hat Pasch mit überzeugender Klarheit wie folgt beschrieben:

> *„Es muß in der Tat, wenn anders die Geometrie wirklich deduktiv sein will, die Deduktion überall unabhängig sein vom Sinn der geometrischen Begriffe, wie er unabhängig sein muß von den Figuren; nur die in den benutzten Sätzen und Definitionen niedergelegten Beziehungen zwischen den geometrischen Begriffen dürfen in Betracht kommen.“* ([16], S. 90.)

Insgesamt ist damit Hilberts Verständnis des Begriffes einer *Mathematischen Theorie* und ihrer axiomatischen Begründung klar umrissen. Die von ihm ausgearbeiteten *Grundlagen der Geometrie* ordnen sich diesem allgemeinen Theorie-Begriff unter.

Literatur

1. Blumenthal, O.: Lebensgeschichte. In: David Hilbert Gesammelte Abhandlungen, Bd. 3, S. 388–429 (1935). Berlin

2. Dedekind, R.: Gesammelte mathematische Werke, 3. Bd. Braunschweig (1932)
3. Einstein, A.: Geometrie und Erfahrung. Sitz. ber. Preuss. Akad. Wiss. Berl. Philos.-Hist. Kl., S. 123–130 (1921)
4. Euklid: Euclidis Elementa, I.L. Heiberg (Hrsg.). Teubner, Leipzig (1883). Deutsche Übersetzung: Die Elemente, Buch I–XIII, herausgegeben und übersetzt von Clemens Thaer. Wiss. Buchgesellschaft, Darmstadt (1962)
5. Felgner, U.: Das Induktionsprinzip. Jahresber. Dtsch. Math.-Ver. **114**, 23–45 (2012)
6. Frege, G.: Kleine Schriften, Herausgegeben Von Ignacio Angelelli. Olms-Verlag, Hildesheim (1967)
7. Freudenthal, H.: Zur Geschichte der Grundlagen der Geometrie, zugleich eine Besprechung der 8. Aufl. von Hilberts „Grundlagen der Geometrie“. Nieuw Archief voor Wiskunde (4), Bd. 5, S. 105–142 (1957)
8. Freudenthal, H.: Die Grundlagen der Geometrie um die Wende des 19. Jahrhunderts. Math.-Phys. Semesterb. **7**, 2–25 (1960)
9. Grattan-Guinness, I.: The Search for Mathematical Roots 1870–1940. Princeton University Press, Princeton (2000)
10. Hallett, M., Majer, U.: David Hilbert's Lectures on the Foundations of Geometry 1891–1902. Springer, Berlin (2004)
11. Hilbert, D.: Grundlagen der Geometrie. Festschrift zur Feier der Enthüllung des Gauss-Weber Denkmals in Göttingen. Teubner, Leipzig (1899) (2. Aufl. 1903, 4. Aufl., 1913, 7. Aufl., 1930, 1968, 14. Aufl., 1999)
12. Hilbert, D.: Über den Zahlbegriff. Jahresber. Dtsch. Math.-Ver. **8**, 180–184 (1900)
13. Husserl, E., Werke, G., Band, X.I.I.: Martinus Nijhoff Verlag. Philosophie der Arithmetik (1890–1901). Martinus Nijhoff Verlag, Den Haag (1970)
14. Iamblichos: De Vita Pythagorica Liber, griechisch und deutsch, herausgegeben, übersetzt und eingeleitet von Michael von Albrecht. Artemis-Verlag, Zürich (1963)
15. Krenek, E.: Über Neue Musik. Verlag der Ringbuchhandlung, Wien (1937)
16. Pasch, M.: Vorlesungen über neuere Geometrie. Teubner, Leipzig (1882). 2. Aufl. Springer, Berlin (1926)
17. Peletier, J. (Peletarius): Euclidis Elementa Geometriae Demonstrationum Libri Sex. Lyon (1557)
18. Poincaré, H.: Les fondements de la Géométrie. Considérations se rapportant à l'ouvrage Grundlagen der Geometrie de Hilbert. Bull. Am. Math. Soc. **10**, 1–23 (1903)
19. Proklos: Procli Diadochi in Primum Euclidis Elementorum Librum Commentarii, Friedlein, G. (Hrsg.). Teubner, Leipzig (1873). Deutsche Übersetzung von P.L. Schönberger: Kommentar zum ersten Buch von Euklids „Elementen“. Halle (Saale) 1945
20. Schmidt, A.: Zu Hilberts Grundlagen der Geometrie. In: Hilberts Gesammelten Abhandlungen, Bd. 2, S. 404–414. Berlin (1933).
21. Schwabhäuser, W., Szmielew, W., Tarski, A.: Metamathematische Methoden in der Geometrie. Springer, Berlin (1983)
22. Tarski, A.: The Completeness of Elementary Algebra and Geometry. Paris (1940). Nachdruck in: Tarskis Collected Papers, Bd. 4, S. 297–346. Birkhäuser, Basel (1986)
23. Tarski, A.: What is elementary geometry. In: Henkin, L., Suppes, P., Tarski, A. (Hrsg.) The Axiomatic Method. North-Holland, Amsterdam (1959). Nachdruck in: Tarskis Collected Papers, Bd. 4, S. 17–32. Birkhäuser, Basel (1986)
24. Toepell, M.-M.: Über die Entstehung von David Hilberts 'Grundlagen der Geometrie'. Vandenhoeck & Ruprecht, Göttingen (1986)
25. Wiener, H.: Über Grundlagen und Aufbau der Geometrie. Jahresber. Dtsch. Math. Ver. **1**, 45–48 (1890/1891). Jahresber. Dtsch Math. Ver. **3**, 70–80 (1892/1893)

Ulrich Felgner geb. 1941, ist Professor (i.R.) für Mathematik an der Universität Tübingen. Er studierte an den Universitäten Gießen, Besançon und Frankfurt/M., promovierte 1968 in Tübingen und habilitierte sich 1973 in Heidelberg. Er wirkte als Professor an den Universitäten Heidelberg, Freiburg und Tübingen. Seine Hauptarbeitsgebiete sind Algebra, Mathematische Logik, Grundlagen und Geschichte der Mathematik.

Jahresber. Dtsch. Math. Ver. (2014) 115:207–209
DOI 10.1365/s13291-013-0075-1

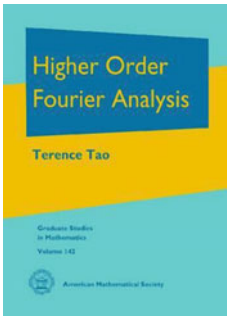
BOOK REVIEW

Terence Tao: "Higher Order Fourier Analysis"

Graduate Studies in Mathematics 142, American Mathematical Society, 2012, 187 pp

Tom Sanders

Published online: 26 November 2013



Classical Fourier analysis concerns itself with the decomposition (where possible) of functions on $\mathbb{Z}$ into linear combinations of (linear) characters, that is functions of the form $z \mapsto \exp(2\pi i\alpha z)$. Decompositions into linear characters are only so useful, and it is increasingly helpful to consider analysis in terms of functions of the form $z \mapsto \exp(2\pi i\alpha z^2)$ and other so-called 'higher order' functions. This analysis is being termed Higher Order Fourier Analysis, and it is the purpose of the book under review to give, to quote the author, a broad tour of this nascent field.

The book is based on a topics graduate course taught by the author in 2010 and is not intended as a substitute for the (rather formidable) core papers on the subject. Instead it focuses on basic foundational material and illustrative examples.

To go into a little more detail, a basic aim with classical Fourier analysis is to decompose a function $f : \mathbb{Z} \to [-1, 1]$ as $f = g + h$ where g is of 'low complexity' (in fact almost periodic) and h is in some sense 'small'. The notion of 'low complexity' can be roughly captured by the idea that g can be written as a linear combination of a small number (if we need a large number we think of the function as having high complexity) of maps $z \mapsto \exp(2\pi i\alpha z)$; to say that h is small we need a suitable norm. A candidate norm is the Gowers U^2-norm defined by

$$\|h\|_{U^2}^4 := \sum_{x,y,z\in\mathbb{Z}} h(x)h(x+y)h(x+z)h(x+y+z).$$

T. Sanders (✉)
Oxford, UK
e-mail: tom.sanders@maths.ox.ac.uk

It turns out to be relatively easy to show that this is a norm and in this case, given $\epsilon \in (0, 1]$ it turns out that we can write $f = g + h$ where g is a linear combination of $O_\epsilon(1)$ maps of the form $z \mapsto \exp(2\pi i \alpha z)$, and $\|h\|_{U^2} \leq \epsilon$.

This fact is at the heart of results from additive number theory such as Roth's theorem (covered in §1.2 of the book) where we are interested in counting three-term arithmetic progressions in large subsets of $\{1, \dots, N\}$. If we write $R_3(A)$ for the number of three-term arithmetic progressions in the set $A \subset \{1, \dots, N\}$ then if B is a set such that $\|1_A - 1_B\|_{U^2}$ is small (in a suitable sense) then $R_3(A) \approx R_3(B)$, and this is a key ingredient in the proof of Roth's theorem. Unfortunately, if we are interested in counting four-term arithmetic progressions then this does not necessarily hold, which is to say that $\|1_A - 1_B\|_{U^2}$ being small does *not* imply $R_4(A) \approx R_4(B)$, where R_4 is the count of four-term arithmetic progressions.

This last fact is one of the reasons for wanting to consider higher order Fourier analysis: one can ask for norms that do have the property that $\|1_A - 1_B\|$ small implies $R_4(A) \approx R_4(B)$. Of course there are many norms for which this is true *e.g.* the ℓ_1-norm, but we need something weaker so that we can hope to have some sort of decomposition of a given function into an 'almost periodic' part and a 'small' part. An appropriate family of norms are the Gowers uniformity norms defined by

$$\|h\|_{U^k}^{2^k} := \sum_{x, x_1, \dots, x_k \in \mathbb{Z}} \prod_{S \subset [k]} h\left(x + \sum_{i \in S} x_i\right).$$

These are developed in §1.3.3 of the book.

Examples of functions with large U^2-norm are given, perhaps unsurprisingly, by maps of the form $z \mapsto \exp(2\pi i \alpha z)$; examples of functions with large U^k-norm are similarly given by maps of the form $z \mapsto \exp(2\pi i p(z))$ where p is a degree $(k-1)$-polynomial. A basic part of understanding higher order Fourier analysis is understanding these examples of functions with large U^k-norm better. §1 of the book deals with the problem of showing that they are equidistributed on the torus. (The case $k = 2$ is trivial but higher values of k are much more demanding.)

One of the recent big advances in this field has been to show that the *only* way that a function can have large U^k-norm is if a large part of it 'looks like' a map of the form $z \mapsto \exp(2\pi i p(z))$ where p is (something like) a degree $(k-1)$-polynomial. This result is called the Gowers Inverse Theorem (due to Green, Tao and Ziegler) and its statement is explained in §1.6 of the book.

The book also covers the inverse conjecture (now theorem) for a different setting called the model setting in §1.5, and includes a sketch of its proof; and, separately, it discusses the most celebrated application of the inverse theorem for the integers, namely, the solving of linear equations in the primes in §1.7.

Finally, the second part of the book deals with some related topics. In particular, the uncertainty principle and its relationship with duality, and also some 'higher order analogues' of Hilbert space resulting from certain multilinear forms in 2^k variables that can be used to define the Gowers uniformity norms.

This area is a technically very demanding area and even understanding the statements of many of the results requires an investment. This book provides a number of very helpful insights from one of the architects of the theory and, as such, is a valuable resource.

Jahresber. Dtsch. Math. Ver. (2014) 115:211–215
DOI 10.1365/s13291-013-0073-3

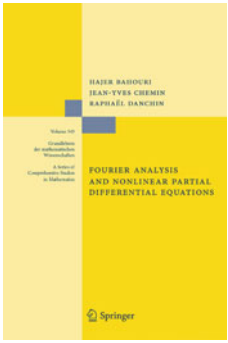
BOOK REVIEW

Hajer Bahouri, Jean-Yves Chemin, Raphaël Danchin: "Fourier Analysis and Nonlinear Partial Differential Equations"

Grundlehren der mathematischen Wissenschaften 343, Springer-Verlag, 2011, 524 pp

Herbert Koch

Published online: 6 December 2013



1 Harmonic Analysis and Partial Differential Equations

Differential operators with constant coefficients commute with translations, and the Fourier transform conjugates those differential operators to the multiplication operator by polynomials. These facts are at the basis of the connection between harmonic analysis and linear and nonlinear partial differential equations.

This is a vast area of mathematics and a comprehensive description is virtually impossible. Instead it may be more appropriate to list and discuss elements and examples.

(1) The Paley–Littlewood decomposition of functions into parts supported in dyadic frequency ranges became an indisponsable tool for differential equations: We fix a partition of unity of the form

$$1 = \sum_{j\in\mathbb{Z}} \phi(2^{-j}\xi)$$

for $\xi \neq 0$. We choose $\phi \in C^\infty(\mathbb{R}^d)$ supported in the annulus $\{\xi : \frac{1}{2} \leq |\xi| \leq 2\}$. It is not difficult to find such a function. Then for any nice function f with Fourier transform $\hat{f}$

$$\hat{f} = \sum_{j=-\infty}^{\infty} \phi(2^{-j}\xi)\hat{f}(\xi)$$

H. Koch (✉)
Bonn, Germany
e-mail: koch@math.uni-bonn.de

is a Paley–Littlewood decomposition of f. It provides a tool to separate the study of solutions to differential equations in different regimes.

(2) For transport Lipschitz regularity of the vector field is a natural condition, but in view of nonlinear problems one wants to study slightly less regular divergence free vector fields. The ordinary differential equation

$$\dot{x} = f(x)$$

has a unique local solution to the initial value problem when f is a Lipschitz continuous vector field. This can be relaxed to a log Lipschitz condition,

$$|f(x) - f(y)| \leq c|x - y| \ln(2 + |x - y|^{-1}).$$

(3) Dispersion is the phenomenon that waves with different frequencies move with different velocities. The prototypical example is the linear Schrödinger equation on $\mathbb{R}^d$

$$i\partial_t u + \Delta u = 0$$

which has the plane wave solutions for $\xi \in \mathbb{R}^d$

$$u(t, x) = e^{i(t|\xi|^2 - x\cdot\xi)}$$

with so called group velocity 2ξ. This leads to pointwise decay for compactly supported initial data: Waves disperse, which can be seen in the pointwise decay of the fundamental solution

$$\left((4\pi it)^{-\frac{1}{2}}\right)^d e^{i\frac{|x|^2}{4t}}.$$

Dispersion is typically connected with a decomposition into wave packets with localized Fourier transforms. On the technical side one is led to oscillatory integrals and to microlocal analysis.

(4) In the context of a Paley Littlewood decomposition nonlinearities are handled via Bony's paradifferential calculus: Denote by $u_{\leq 2^k}$ the inverse Fourier transform of

$$\sum_{j\leq k} \phi(2^{-j}\xi)\hat{u}(\xi)$$

and by u_{2^k} the inverse Fourier transform of

$$\phi(2^{-k}\xi)\hat{u}(\xi)$$

Then one has for smooth functions f and nice functions u

$$\begin{aligned} f(u) &= \lim_{k\to\infty} f(u_{\leq 2^k}) \\ &= \sum_k f(u_{\leq 2^k}) - f(u_{\leq 2^{k-1}}) \\ &= \sum_k \int_0^1 f'\left(u_{\leq 2^{k-1}} + s u_{2^k}\right) ds u_{2^k}. \end{aligned}$$

This formula relates nonlinear functions and the Paley-Littlewood decomposition.

The book at hand provides a coherent selection of topics belonging to these points where the selection is due to the interest of the authors. The important and relevant problems covered in this book display a fascinating interplay between transport, dispersion and nonlinear behaviour, a highly active and interesting field. Most of the problems covered have some scaling symmetry, which singles out function spaces invariant under this scaling. They are called critical. To be specific, consider the incompressible Navier–Stokes equations on $(0, \infty) \times \mathbb{R}^d$,

$$u_t + u\nabla u - \Delta u + \nabla p = 0$$

where u is a divergence free velocity vector field, and p is the pressure. $(u\nabla u)^l = \sum_{j=1}^d (u^j \partial_j) u^l$, and Δu is the Laplacian on the components. The acceleration of a fluid particle is given by $u_t + u\nabla u$. The flow is driven by the pressure p, and damped by the viscosity through Δu. The pressure itself occurs as Lagrangian multiplier related to the incompressibility of the flow resp. divergence freeness of the vector field.

If u and p satisfy the Navier–Stokes equations and $\lambda > 0$ then so do the functions $\lambda u(\lambda^2 t, \lambda x)$ and $\lambda^2 p(\lambda^2 t, \lambda x)$. The Lebesgue space $L^q(\mathbb{R}^d)$ is invariant under this scaling exactly when $q = d$, i.e.

$$\int_{\mathbb{R}^d} |\lambda u(0, \lambda x)|^d dx = \int_{\mathbb{R}^d} |u(0, x)|^d dx.$$

It is called critical. A major focus of the authors is on problems in critical spaces.

2 Contents

The first two chapters introduce background information and collect results used later. The discussion of partial differential equations begins in Chapter 3 with linear transport and transport-diffusion equations. Characteristic for this part of the book is the study of transport equations in Besov spaces and of Log-Lipschitz vector fields. Chapter 4 introduces quasilinear symmetric systems, a highly relevant class of hyperbolic evolution equations. As in the rest of the book the focus is on strong resp. mild solutions, but not on weak solutions with shocks or rarefraction waves. On the other hand

the authors carefully push the treatment of classical solutions to the limit, and discuss the flow map and critical spaces.

The incompressible Navier–Stokes equations are the object of study of chapter 5. The chapter provides a fairly complete study of 'mild' solutions, with an emphasis again on critical spaces. It includes global existence for small data in critical spaces like $\dot{H}^{d/2-1}$, $L^3(\mathbb{R}^3)$, $\dot{B}^{-1+n/p}_{p,\infty}$ and BMO^{-1}, a study of the set of initial data with global solutions, and a study of the flow map defined by the velocity field of the solution.

In geophysics there are vastly different vertical and horizontal scales. This motivates to consider the anisotropic 3d Navier–Stokes equations

$$u_t + (u\nabla)u - (\partial_1^2 + \partial_2^2)u + \nabla p = 0$$

$$\nabla \cdot u = 0$$

in chapter 6, for which the equation for the vertical (third) component is of hyperbolic nature. The authors study this problem again in the scale invariant spaces $H^{0,1/2}$ where half a vertical derivative is in L^2, and a variant of it.

Here the contraction argument breaks down, similar to what happens for quasilinear hyperbolic systems, and the standard strategy of regularizing, a priori estimates, passage to the limit und uniqueness estimates in weak norms provides the proof. While the strategy is standard its implementation is not, and it is fairly recent.

The same strategy is applied in Chapter 7 to the Euler equations of incompressible inviscid fluids. The main result states local wellposedness in Besov spaces which imbed into C^1 and global existence in dimension 2. The characterization of the maximal existence interval due to Kato and Majda, Yudovich's existence result for bounded vorticity in 2d and the inviscid limit are the highlights of the first part of this chapter. It concludes with the striking result that the regularity of vortex patches is preserved under the flow—an unexpected result due to Chemin, which is extended to a study of the inviscid limit with 'striated' regularity.

Chapter 8 studies the nonlinear Schrödinger equation and the nonlinear wave equation via (refined) Strichartz estimates, and the contraction mapping principles. It includes a proof of the endpoint Strichartz estimates. The considered nonlinearity is u^p for suitable p. The chapter culminates in global existence for the defocussing cubic nonlinear wave in 3 space dimensions below energy in $\dot{H}^{3/4} \times \dot{H}^{-1/4}$, a study initiated in Bourgain's work in two space dimensions.

The quasilinear wave equation

$$u_{tt} - \Delta u - \sum_{i,j=1}^{d} \partial_i (g_{ij}(u)\partial_j u) = \sum_{i,j=1}^{d} Q_{ij}(u)\partial_i u \partial_j u$$

is in the focus of Chapter 9. In slightly simplified formulation the main result is local wellposedness for $u(0,.) \in H^{\frac{d}{2}+\frac{3}{4}}$ and $u_t(0,.) \in H^{\frac{d}{2}-\frac{1}{4}}$ if $d \geq 4$, with similar statements for $d = 2$ and $d = 3$. Here H^s denotes the Sobolev spaces of functions with s derivatives in L^2, where s is allowed to be a real number.

The study of compressible barotropic Navier–Stokes equations combines many of the techniques and phenomena of the whole book. It is a mixed hyperbolic-parabolic system. Following papers of the third author local existence for large data and global existence for small deviation from the trivial constant density solution, and convergence to the incompressible Navier–Stokes equations is proven. The proof requires dispersive techniques to control acoustic dispersive waves.

3 Discussion

The authors did make impressive contributions to a broad area of fluid dynamics. It is the first time that a coherent presentation of those research results is available, which will give easier access to the whole area to a broader audience. This is a very valuable addition to the existing literature, written by leading mathematicians who have been involved in developing the whole theory.

No book can cover an vast area indicated by the title. The authors focus on areas which they and their collaborators have shaped—even if the title of the book and some of the sections suggest a broader scope. It is a continuation of the books Chemin [3] and Chemin et al. [2], with little overlap with the books by Bourgain [1] and Tao [4] on dispersive equations.

The book is fairly selfcontained and complete, starting from topics like duality in L^p. Whether an undergraduate understanding of analysis suffices—as stated in the introduction—for its understanding remains to be seen.

It is a valuable contribution in the important area of the interest of the authors and will without question find its place in the mathematical libraries, and on the shelves of people working in those areas.

References

1. Bourgain, J.: Global Solutions of Nonlinear Schrödinger Equations. American Mathematical Society Colloquium Publications, vol. 46. American Mathematical Society, Providence (1999)
2. Chemin, J.-Y., Desjardins, B., Gallagher, I., Grenier, E.: Mathematical Geophysics. An Introduction to Rotating Fluids and the Navier–Stokes Equations. Oxford Lecture Series in Mathematics and its Applications, vol. 32. The Clarendon Press, Oxford University Press, Oxford (2006)
3. Chemin, J.-Y.: Perfect Incompressible Fluids. Oxford Lecture Series in Mathematics and its Applications, vol. 14. The Clarendon Press, Oxford University Press, New York (1998) (Translated from the 1995 French original by Isabelle Gallagher and Dragos Iftimie)
4. Tao, T.: Nonlinear Dispersive Equations. Local and Global Analysis. CBMS Regional Conference Series in Mathematics, vol. 106. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence (2006)

Jahresber Dtsch Math-Ver (2014) 115:217–221
DOI 10.1365/s13291-013-0070-6

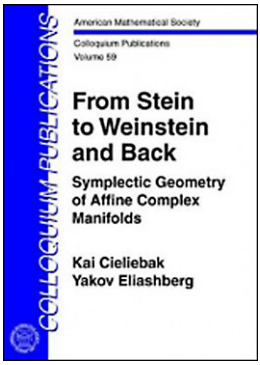
BOOK REVIEW

Kai Cieliebak and Yakov Eliashberg: "From Stein to Weinstein and Back. Symplectic Geometry of Affine Complex Manifolds"

AMS, 2012, 364 pp

Felix Schlenk

Published online: 28 August 2013



Complex geometry is an old and rich field with its origin in the study of complex functions (which in the nineteenth century were considered as much more natural than real functions, for instance because complex polynomials always decompose into linear factors). Stein manifolds are affine complex manifolds. Classically, such manifolds are studied up to biholomorphism. This book is not concerned with this fine structure of Stein manifolds, but with their topological structure. The approach is through symplectic geometry. Symplectic geometry is a much younger field, with its roots in classical mechanics and Hamiltonian systems, where the phase spaces carry a canonical symplectic structure. Since Gromov's introduction of J-holomorphic methods and Floer's creation of his homology, symplectic geometry is linked to many other fields (complex analysis, algebraic geometry, mathematical physics, low-dimensional topology, etc.) with cross-fertilization in both directions. This book demonstrates that symplectic geometry is also the structure that governs the topology of Stein manifolds. It is about the symplectic geometry of Stein manifolds and its implications for the complex geometry of Stein manifolds.

In the sequel, I first explain the main notions (Stein and Weinstein manifolds), then describe the main results proven in this book, and finally say something about the style of exposition and how this book came to life.

F. Schlenk (✉)
Neuchâtel, Switzerland
e-mail: felix.schlenk@unine.ch

Stein Manifolds. A *complex manifold* is a smooth manifold V with an atlas for which all transition functions are biholomorphic mappings. The complex multiplication i on $\mathbb{C}^n$ and the charts for V then induce the *complex structure* J on the tangent bundle TV. A complex embedding of a complex manifold (V, J) into some complex vector space $(\mathbb{C}^N, i)$ is an embedding such that the complex structure J on $V \subset \mathbb{C}^N$ becomes the restriction of i to TV. While every smooth manifold can be embedded into some real vector space $\mathbb{R}^N$, not every complex manifold admits a complex embedding into some complex vector space $(\mathbb{C}^N, i)$. Indeed, if V is a complex manifold that is compact without boundary, then this is impossible in view of the maximum principle. Complex manifolds that admit proper complex embeddings into some $\mathbb{C}^N$ are called *Stein manifolds*. Stein manifolds are thus the "affine complex manifolds". They were first considered by Karl Stein in 1951 and named after Stein by H. Cartan.

Examples of Stein manifolds are $\mathbb{C}^n$, closed Riemann surfaces deprived from at least one point, and, more generally, the complement $\mathbb{C}\mathrm{P}^n \setminus H$ of any hyperplane H of a complex projective space.

Every Stein manifold (V, J) is a symplectic manifold. Indeed, every complex embedding into $\mathbb{C}^N$ provides a symplectic form on V by restricting to V the usual symplectic form $\omega_{\mathrm{st}} = \sum_{j=1}^N dx_j \wedge dy_j$ of $\mathbb{C}^N$. To see that this symplectic structure does not depend on the embedding, it is useful to characterize Stein manifolds intrinsically: To a smooth function $\phi\colon V \to \mathbb{R}$ associate the 1-form $d^{\mathbb{C}}\phi := d\phi \circ J$ and the 2-form $\omega_\phi := -dd^{\mathbb{C}}\phi$. The function ϕ is called *J-convex* (or *strictly plurisubharmonic*) if $g_\phi(v, w) := \omega_\phi(v, Jw)$ defines a Riemannian metric on V. In particular, the 2-form ω_ϕ is symplectic (i.e., closed and non-degenerate). A function $\phi\colon V \to \mathbb{R}$ is called *exhausting* if it is proper (i.e., preimages of compact sets are compact) and bounded from below.

The function $\phi_{\mathrm{st}}(z) := |z|^2$ on $\mathbb{C}^N$ is exhausting and i-convex, with $\omega_\phi = \omega_{\mathrm{st}}$. Hence every Stein manifold (V, J) admits an exhausting J-convex function. The converse is also true according to results of Grauert, Bishop and Narasimhan: Every complex manifold (V, J) admitting an exhausting J-convex function ϕ is Stein. It is now not too hard to show that given two exhausting J-convex functions ϕ, ψ on (V, J), the two symplectic forms ω_ϕ, ω_ψ are diffeomorphic. It follows that Stein manifolds carry a canonical symplectic structure.

The space of J-convex functions on V is contractible and C^2-open. A generic J-convex function is thus Morse (i.e., all critical points are non-degenerate), and a generic path of J-convex functions consists of *generalized Morse functions*, i.e., functions ϕ_t that near a degenerate critical point p of, say, ϕ_0 in appropriate coordinates $x_1, \ldots, x_m$ look like the birth–death family

$$\phi_t(x) = \phi_t(p) \pm t x_1 + x_1^3 - \sum_{i=2}^{k} x_i^2 + \sum_{j=k+1}^{m} x_j^2.$$

A *Stein structure* on V is a pair (J, ϕ) where J is a complex structure on V and ϕ is a J-convex generalized Morse function. The Stein structures corresponding to two exhausting J-convex functions on the same complex manifold are homotopic. In particular, the space $\mathfrak{Stein}(V)$ of Stein structures on V is connected. The main topic of this book is the study of the topology of the spaces $\mathfrak{Stein}(V)$. To this end, the underlying symplectic geometry of V is formalized:

Weinstein Manifolds. A *Weinstein structure* on an open manifold V is a triple (ω, X, ϕ), where

- ω is a symplectic form on V,
- $\phi\colon V \to \mathbb{R}$ is an exhausting generalized Morse function,
- X is a complete vector field which is Liouville for ω and gradient-like for ϕ.

Here, X is Liouville for ω if its Lie derivative preserves ω, i.e., $\iota_X \omega = \omega$. The quadruple (V, ω, X, ϕ) is then called a *Weinstein manifold*. Such manifolds were introduced by Alan Weinstein in 1991 and named after Weinstein by Eliashberg and Gromov. The space $\mathfrak{Weinstein}(V)$ of Weinstein structures on V is also connected. Examples of Weinstein manifolds are:

(1) $\mathbb{R}^{2n}$ with the "radial" Weinstein structure

$$\omega_{\mathrm{st}} = \sum_j dx_j \wedge dy_j, \qquad X_{\mathrm{st}} = \frac{1}{2}\sum_j \left(x_j \frac{\partial}{\partial x_j} + y_j \frac{\partial}{\partial y_j} \right), \qquad \phi_{\mathrm{st}} = \sum_j (x_j^2 + y_j^2);$$

(2) the cotangent bundle T^*Q over a closed base manifold Q, with the "fibrewise radial" Weinstein structure

$$\omega_{\mathrm{st}} = \sum_j dp_j \wedge dq_j, \qquad X = \sum_j p_j \frac{\partial}{\partial p_j}, \qquad \phi = \sum_j p_j^2$$

(this is not quite a Weinstein structure since $\sum_j p_j^2$ is not a generalized Morse function, but this structure can be perturbed to a genuine Weinstein structure).

In a Weinstein manifold (V, ω, X, ϕ), there is an intriguing interplay between Morse theoretic properties of ϕ and symplectic geometry. For instance, the stable manifold W_p^- (with respect to the flow of X) of a critical point p is *isotropic* (meaning that ω vanishes along W_p^-). In particular, the Morse indices of ϕ are $\leq n = \frac{1}{2}\dim V$.

From Stein to Weinstein and Back. Given a Stein structure (J, ϕ), define again the symplectic form $\omega_\phi = -dd^{\mathbb{C}}\phi$. Then the vector field X_ϕ defined by $\iota_{X_\phi}\omega_\phi = -d^{\mathbb{C}}\phi$ is Liouville and gradient-like with respect to the Riemannian metric $g_\phi := \omega_\phi(\cdot, J\cdot)$. It may not be complete, but becomes complete if we compose ϕ with an appropriate function on $\mathbb{R}$. Consider now the map

$$\mathfrak{W}\colon \mathfrak{Stein} \to \mathfrak{Weinstein}, \qquad (J, \phi) \mapsto (\omega_\phi, X_\phi, \phi).$$

This map forgets the most rigid structure (the complex structure) and only retains the far more flexible Weinstein structure (ω_ϕ, X_ϕ). The first important result of this book is

Theorem 1 *The map* $\mathfrak{W}\colon \mathfrak{Stein} \to \mathfrak{Weinstein}$ *induces an isomorphism on* π_0 *and a surjection on* π_1.

In particular, the Stein structure of a Stein manifold can be recovered from its underlying Weinstein structure, uniquely up to homotopy. Theorem 1 and a few other

results in the book lend evidence to the conjecture that $\mathfrak{W}\colon \mathfrak{Stein} \to \mathfrak{Weinstein}$ is a homotopy equivalence. There is hope that this conjecture can be proven by further developing the techniques provided in this book.

Existence of Stein Structures. Theorem 1 also says that every Weinstein manifold carries a Stein structure, unique up to homotopy. The complex-geometric problem of classifying Stein structures (up to homotopy) is thus translated to the symplecto-geometric problem of classifying Weinstein structures (up to homotopy).

For a given smooth manifold V of dimension $2n$, a necessary condition to carry a Weinstein structure is that the tangent bundle TV carries an *almost complex structure* J, i.e., an endomorphism satisfying $J^2 = -\mathrm{Id}$ on each fibre. This is a topological condition on TV that can be understood in terms of obstruction theory. (For instance, the odd Stiefel–Whitney classes of TV must vanish and the even ones must have integral lifts.) As we have already noticed, the Morse indices of the critical points of ϕ cannot exceed n. By Morse theory it thus follows that V has a handlebody decomposition of handles of dimension $\leq n$. In particular, all homology groups $H_i(V;\mathbb{Z})$ with $i > n$ must vanish. Surprisingly, if $2n \neq 4$, these two conditions are already sufficient for the existence of a Weinstein structure, and hence, by Theorem 1, of a Stein structure:

Theorem 2 *A smooth manifold V of real dimension $2n \neq 4$ admits a Stein structure if and only if it admits an almost complex structure and an exhausting Morse function without critical points of index $> n$.*

The main tools in the proof (of both theorems) are suitable h-principles and Morse theory for J-convex functions. In particular, the Morse-theoretic operations that are used in the proof of the h-cobordism theorem—reordering of critical points, handle sliding, and cancellation of critical points—are performed in the class of J-convex functions. The theorem fails for $n = 2$. For instance, $S^2 \times \mathbb{R}^2$ (which of course admits a complex structure) does not admit a Stein structure.

Outlook. This book contains many other important results that have been obtained by the two authors over the last years. We refer to the book for these findings. We would like to mention, however, that the results in this book have already led to spectacular progress in symplectic, contact and complex geometry. We describe one application to each of these geometries, that have all been proved over the last two years.

(1) Sylvain Courte [2] showed that there exist contact manifolds that are not diffeomorphic as smooth manifolds but have symplectomorphic symplectizations.
(2) In dimension ≥ 5, Seidel and Smith, McLean and others had constructed exotic Stein structures on the smooth ball with exotic contact structures on the boundary. Murphy, Niederkrüger, Plamenevskaya and Stipsicz [4] recently showed that when summed with certain overtwisted structures on the sphere, this type of exotic structures immediately disappear. (Indeed, the exotic structure becomes the overtwisted structure.)

(3) Cieliebak and Eliashberg [1] gave a topological characterization of polynomially and rationally convex domains in $\mathbb{C}^n$, $n \geq 3$.

We expect many other important applications, in particular to higher dimensional contact topology, a field in full thrive.

Style of Exposition. This book is a remarkable mix of classical topics and research done by the authors over the last twenty years. It is both a textbook and a research monograph. While all the novel constructions are given in great detail, more classical material, such as topics from complex analysis, symplectic and contact geometry, differential topology (Morse theory, h-cobordism, h-principles), are presented from the perspective of their applications in the book. It is fascinating and refreshing to see these classical tools in action, combined and applied to create something new. The exposition is very beautiful; whenever possible, the geometry of the situation is fully brought out, and formulas and computations are used only to check a geometric intuition.

How the Book Came to Life. In 1990, Yasha Eliashberg [3] published a short paper, proving Theorem 2. In 1996, he gave a Nachdiplomvorlesung (a 28 hours graduate course) at ETH Zürich, with the goal of giving a more detailed account of the proof of Theorem 2. Kai Cieliebak, a finishing graduate student, took notes, and I was a beginning graduate student. While the extremely charismatic lecturer and the breadth of the manifold theories coming up made a strong impression on me and convinced everyone that this must be great mathematics, it was hard to follow the exposition, that remained incomplete. The present book not only contains a detailed proof of Theorem 2, but proves many more theorems, that open up a whole new world at the crossroad of complex, symplectic and contact geometry.

References

1. Cieliebak, K., Eliashberg, Ya.: The topology of rationally and polynomially convex domains. arXiv:1305.1614
2. Courte, S.: Contact manifolds with symplectomorphic symplectizations. arXiv:1212.5618
3. Eliashberg, Ya.: Topological characterization of Stein manifolds of dimension > 2. Int. J. Math. **1**, 29–46 (1990)
4. Murphy, E., Niederkrüger, K., Plamenevskaya, O., Stipsicz, A.: Loose Legendrians and the plastikstufe. arXiv:1211.3895

Jahresber. Dtsch. Math. Ver. (2014) 115:223–228
DOI 10.1365/s13291-013-0074-2

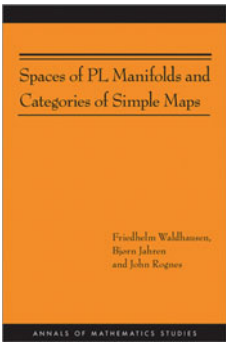
BOOK REVIEW

Friedhelm Waldhausen, Bjørn Jahren, John Rognes: "Spaces of PL Manifolds and Categories of Simple Maps"

Annals of Mathematics Studies 186, Princeton University Press, 2013, 192 pp

Michael Weiss

Published online: 3 December 2013



The book by Waldhausen, Jahren and Rognes is about the approximation, simulation and substitution of homeomorphisms between manifolds by maps which are combinatorially defined. In the setting of the book, the manifolds themselves are typically equipped with a piecewise linear (PL) structure, that is, an atlas with changes of charts which are PL. The homeomorphisms are also required to be PL. This should not be seen as a severe restriction and it does not make the approximation and simulation task trivial. For one thing, there are standard ways to make a space out of the PL homeomorphisms from one PL manifold M to another, N, and it is well understood how the homotopy properties of that space differ from those of the space of all homeomorphisms from M to N. The difference is small and well understood by [8]. For another thing, the technical purpose of the approximations and substitutions in this book is to make a passage from the manifold setting to one without manifold assumptions, and to obtain a dimension-independent model.

The book provides proofs and details for something which took shape more than thirty years ago as forward references in a sequence of articles by Waldhausen, e.g. [18–21], on algebraic K-theory and spaces of h-cobordisms. Preliminary versions of major parts of the book have been in circulation for many years, but not in wide circulation. Apart from the introduction and the introductory chapter 1, the book is fairly elementary and self-contained, and does not mention algebraic K-theory. Chapter 1 explains patiently what the connection with algebraic K-theory is and where the forward references appeared, but the explanations are for readers with a

M. Weiss (✉)
Münster, Germany
e-mail: m.weiss@uni-muenster.de

working knowledge of algebraic topology. This review is also an attempt to describe the context of the book at a more basic level. It is possible to write endlessly about the history of the topic. I have tried to exercise restraint but I may have failed.

The concept of an h-cobordism goes back to Smale's h-cobordism theorem [16,13]. The relationship to algebraic K-theory emerged shortly after that with a version of the h-cobordism theorem for nonsimply-connected manifolds, the s-cobordism theorem of Barden-Mazur-Stallings [12,17,6]. An h-cobordism on a closed (= compact with empty boundary) m-manifold M is a compact manifold W of dimension $m+1$ whose boundary is identified with a topological disjoint union $M \coprod M'$ of closed m-manifolds in such a way that the "inclusions" $M \to W$ and $M' \to W$ are homotopy equivalences. Smale used Morse theory to show, in the differentiable manifold setting, that such an h-cobordism is always diffeomorphic to a product $M \times [0,1]$ if M is simply connected and $m \geq 5$. Dropping the assumption of simple-connectedness, Barden, Mazur and Stallings were able to classify the h-cobordisms on a closed connected M of dimension ≥ 5 up to diffeomorphism or PL-homeomorphism, relative to M. The classes turned out to be in bijection with the elements of $\mathrm{Wh}(\pi_1 M)$, the Whitehead group of the fundamental group $\pi_1 M$. The Whitehead group is a direct summand (usually defined as a quotient) of the algebraic K-group $K_1(\mathbb{Z}[\pi_1 M])$ of the group ring $\mathbb{Z}[\pi_1 M]$; the complementary summand is isomorphic to the abelianization of $\pi_1 M$ times $\mathbb{Z}/2$. And $K_1(R)$, for a ring R, is the abelianization of the direct limit, $n \to \infty$, of the groups $\mathrm{GL}_n(R)$.

The Whitehead group was introduced by Whitehead [24,25] in the course of a very original investigation on the practical meaning and construction of homotopy equivalences. Whitehead called two finite simplicial complexes X and Y *simple homotopy equivalent* if, up to isomorphism, Y could be obtained from X by a finite sequence of elementary collapses and elementary expansions. An elementary collapse is the removal of two simplices σ, τ in dimensions k and $k+1$, where σ is a face of τ but not a face of any other $(k+1)$-simplex; an elementary expansion is the same in reverse. Simple homotopy equivalence is sharper than homotopy equivalence, and the Whitehead group accounts for the difference. Specifically an inclusion $X \to Y$ of simplicial complexes (connected X for simplicity) which is also a homotopy equivalence determines a Whitehead torsion, an element in the Whitehead group of $\pi_1 X$. The Whitehead torsion is zero if and only if Y can be reduced to X by elementary expansions and collapses which do not involve simplices in X. All elements of the Whitehead group of $\pi_1 X$ arise in this way. This was what Barden, Mazur and Stallings used in their classification of h-cobordisms in the PL setting; the classification was given by the Whitehead torsion of the inclusion $M \to W$, for an h-cobordism W on M. In particular the h-cobordisms with zero Whitehead torsion are equivalent to a product $M \times [0,1]$. These are called s-cobordisms, s for *simple homotopy equivalence*. Much of that is beautifully explained or outlined in [2,14].

In the 1960s, h-cobordisms became more important in manifold topology as surgery theory took off. A key task in surgery theory is to classify closed m-manifolds N which are (simple) homotopy equivalent to a fixed closed m-manifold M. Surgery reduces this problem to a mix of algebraic topology and difficult algebra of quadratic forms when $m \geq 5$. More precisely, we can make a "moduli space" $\mathcal{S}(M)$ out of pairs (N, f) where N is a closed m-manifold and $f : N \to M$ is a (simple) homotopy equivalence.

Surgery theory is very successful in describing the set of path components $\pi_0\mathcal{S}(M)$. It is not entirely successful on its own in describing the higher homotopy groups, let alone the homotopy type, of $\mathcal{S}(M)$. But as Hatcher pointed out in [5], it is successful in reducing questions about the homotopy type of $\mathcal{S}(M)$ to questions about the homotopy types of certain spaces of h-cobordisms. This was later made more explicit in [22]. These are not just h-cobordisms on M but h-cobordisms on $M \times D^k$ for all $k \geq 0$. Here two remarks are in order. Firstly, the space $H(M)$ of h-cobordisms on M should be imagined as a space equipped with a bundle $E \to H(M)$ whose fibers are h-cobordisms on M and which is universal among such bundles. This characterizes $H(M)$ up to homotopy equivalence. Secondly, it is possible to make some sense of $H(M)$ when M is a compact manifold with nonempty boundary, although I will not take the time to do so. Indeed $M \times D^k$ for $k > 0$ will have nonempty boundary. The h-cobordisms are meant to be differentiable or PL, etc., depending on the version of $\mathcal{S}(M)$ to be investigated.

There are maps

$$H(M \times D^k) \to H(M \times D^{k+1}) \tag{1}$$

given by taking an h-cobordism W on $M \times D^k$ to the product $W \times [0, 1]$, viewed as an h-cobordism on $M \times D^k \times [0, 1] \cong M \times D^{k+1}$. Hatcher's analysis of the homotopy group $\pi_r\mathcal{S}(M)$ relying on surgery theory and h-cobordism spaces is much more transparent in the case where the stabilization maps (1) are all $(r+1)$-connected. Consequently this became an important question in manifold topology: are the maps (1) highly connected?

Meanwhile algebraic K-theory had also grown up by 1970. Among many definitions of the algebraic K-theory space $K(R)$ for a ring R, I find the following, taken from [1], very illuminating although it is not very useful in computations. In the very beginning of K-theory there was the Grothendieck construction which turns abelian monoids into abelian groups. In homotopy theory there is a construction with a similar purpose, called group completion, which turns associative topological monoids into grouplike (associative) topological monoids. Here *grouplike* means that the multiplication admits an operation of inverse, at least up to homotopy. Let B_R be the space which carries the universal bundle of finitely generated projective left R-modules. It has a multiplication $B_R \times B_R \to B_R$, sufficiently associative, which reflects the operation of fiberwise direct sum for bundles of finitely generated projective left R-modules. Apply group completion to B_R . The result is $K(R)$. The group of path components of $K(R)$ is $K_0(R)$, the projective class group. The fundamental group of $K(R)$ is what was known before as $K_1(R)$. The higher K-groups $K_i(R)$ are by definition the homotopy groups $\pi_i K(R)$, for $i > 1$.

Now, since for a compact PL manifold M of dimension ≥ 5 the set of path components of the h-cobordism space $H(M)$ has been identified with $\mathrm{Wh}(\pi_1 M)$, a direct summand of $K_1(\mathbb{Z}\pi_1 M)$, it can be asked whether the space $H(M)$ as a whole, in the PL setting or any other, admits a description in algebraic K-theory terms. This became another important question in manifold topology.

The two important questions were taken on by Hatcher and Waldhausen respectively in the mid 1970s. Probably I am simplifying excessively, but it seems to me that

Hatcher concentrated on the first problem, trying to show that the maps (1) are highly connected. Waldhausen concentrated on the second problem, making the connection between h-cobordism spaces and algebraic K-theory, but taking care to circumvent the first problem. Hatcher [4] introduced a few fundamental ideas into the subject. He associated with a space X (described in combinatorial terms, say as a simplicial complex) another space, the *Whitehead space* of X, which was intended as a manifold-free and dimension-independent simulation of the h-cobordism space of X in the case of a compact PL manifold X. This was broadly adopted by Waldhausen, e.g. in [18], but a few details were changed in the light of rumors that Hatcher's definition of the Whitehead space of X was not as robust as advertised. Also it was noted that Hatcher's Whitehead space was the loop space (space of pointed maps from S^1 to something) of another pointed space which, as Hatcher and Waldhausen seemed to agree at this point, was even more deserving of the title *Whitehead space of X* although it does not need to concern us here. Consequently it is easier for me to sketch the definition of Hatcher's [4] of Whitehead space after the Waldhausen improvements, and it is my duty to denote it by $\Omega\mathrm{Wh}(X)$ or $\Omega\mathrm{Wh}^{\mathrm{PL}}(X)$, where the Ω is for *loop space*. Assume that X is a simplicial set, and for simplicity assume also that it is a finitely generated simplicial set. (Finitely generated simplicial sets are a mild variation on finite simplicial complexes.) Waldhausen makes a category whose objects are injective maps $X \to Y$ of finitely generated simplicial sets, fixed X and variable Y, which turn into homotopy equivalences $|X| \to |Y|$ after geometric realization (indicated by the vertical lines). A morphism from $(X \to Y)$ to $(X \to Y')$ is a map of simplicial sets $f : Y \to Y'$ which is *simple* (which means that, after geometric realization, all its point inverses are contractible) and respects the reference maps from X. The space $\Omega\mathrm{Wh}(X)$ is the classifying space of this category [15]. Note the overwhelming similarity with the geometric description of the Whitehead group $\mathrm{Wh}(\pi_1 X)$. The simple maps in the definition of $\Omega\mathrm{Wh}(X)$ are inspired by Whitehead's elementary collapses. (A simple map is a simple homotopy equivalence, a not-completely-trivial fact which others had noticed before. Not all simple homotopy equivalences are simple maps.) Related to that there is a bijection of $\pi_0\Omega\mathrm{Wh}(X)$ with the Whitehead group $\mathrm{Wh}(\pi_1 X)$. The definition of $\Omega\mathrm{Wh}(X)$ is functorial in that a map $f : X \to X'$ of simplicial sets induces a map $f_* : \Omega\mathrm{Wh}(X) \to \Omega\mathrm{Wh}(X')$. If f is a homotopy equivalence after geometric realization, then f_* is a homotopy equivalence from $\Omega\mathrm{Wh}(X)$ to $\Omega\mathrm{Wh}(X')$.

In the case where $X = M$ is a compact PL manifold, both Hatcher and Waldhausen have a comparison map from $H(M)$, in the PL version, to $\Omega\mathrm{Wh}(M)$. This is what I have called the substitution or simulation, *la raison d'être* for much of the book by Waldhausen, Jahren and Rognes. If I understand correctly, it is difficult to set this up. (In case the connection with PL homeomorphisms is unclear, recall that $H(M)$ carries a universal bundle of h-cobordisms, and bundles tend to have structure groups. Here the structure groups are groups of PL homeomorphisms.) Hatcher attempted to show that the comparison map is highly connected; the proposed lower bound on connectivity was approximately a third of the dimension of M. That would have implied that (1) is highly-connected, too. Regardless of that, the homotopy invariance and naturality properties of the functor $X \mapsto \Omega\mathrm{Wh}(X)$ imply that the comparison map $H(M) \to \Omega\mathrm{Wh}(M)$ extends to a map

$$\mathscr{H}(M) \longrightarrow \Omega \mathrm{Wh}(M) \tag{2}$$

where $\mathscr{H}(M) = \mathrm{colim}_k H(M \times D^k)$ is the direct limit of the h-cobordism spaces $H(M \times D^k)$ under the maps (1). One of the main results of the book, I would say *the* main result, is that (2) is a homotopy equivalence, with the PL interpretation of $\mathscr{H}(M)$. With that result in place, making the connection between spaces of h-cobordisms and algebraic K-theory amounts to making the connection between $\Omega \mathrm{Wh}(X)$ and algebraic K-theory. That was actually done long ago in [18], with only very minor forward references to the book under review. There is also a variant for the differentiable setting, and another for the plain topological setting; they are formulated in the book, and are proved by reduction to the PL setting. These reductions also go back to papers by Waldhausen written long ago, especially [19,20]. I only want to emphasize here that $\mathscr{H}(M)$ in the differentiable setting is usually not homotopy equivalent to $\mathscr{H}(M)$ in the PL setting, even though the experience of the s-cobordism theorem seems to suggest that the two are homotopy equivalent.

The idea of a *simple map* (map with contractible point inverses) is much older than this book and the related articles by Waldhausen. J.H.C. Whitehead had some elementary forms of it, Cohen [3] had PL maps between simplicial complexes with contractible point inverses which he called *contractible mappings*, and Lacher wrote influential papers about cell-like maps [9–11] which are non-combinatorial analogues. There were important results about the approximation of simple maps (and variants) by homeomorphisms, by Cohen in the PL case and by Siebenmann in the setting of topological manifolds. But the book by Waldhausen, Jahren and Rognes is a unique and valuable resource on the topic of simple maps between combinatorial objects (such as finitely generated simplicial sets) and should remain a standard reference for many years.

An important theme of the book which I have mentioned too little until now is desingularization, the art of approximating simplicial sets by, e.g., simplicial complexes via simple maps. Apart from the introductory chapter 1 the book has chapters 2, 3 and 4. Chapter 2 has a lot of material on desingularization. Here the authors have done enormous work in digging out and breathing new life into long-forgotten treatises. Let us note that the category of simplicial sets, in contrast with the category of simplicial complexes, shares many good properties with the category of sets, such as, existence of all categorical inverse and direct limits. Waldhausen's decision to use simplicial sets in the definition of $\Omega \mathrm{Wh}(X)$ meant that the connection from there to algebraic K-theory was easier to make, but we learn by reading this book that it had a price. Chapter 3 of the book deals with the transition from categories of simple maps with morphism *sets* to (enriched) categories of simple maps with morphism *spaces*. It is only in chapter 4 that manifolds come into the picture.

The book has been written with enormous patience but it is not for impatient readers. For me, appreciation came gradually with meditations on the vast historical context of the book and the fascinating pitfalls of combinatorial topology.

The problem of deciding whether the maps (1) are highly connected and of giving useful estimates for that remains open in the PL setting, as Waldhausen, Jahren and Rognes point out in their book. This is very curious, and one cannot help wondering whether a very very careful reading of their book might take us closer to an answer. In the differentiable setting, Igusa [7] gave a solution. Another problem in the area which

I find tantalizing is as follows. Let $Z_k(M)$ be the homotopy fiber of the stabilization map (1), either in the PL setting or in the differentiable setting. (The notation is obviously provisional.) There are stabilization maps $Z_k(M) \to \Omega Z_{k+1}(M)$ analogous to the stabilization maps (1). The direct limit $\operatorname{colim}_k \Omega^k Z_k(M)$ can be formed, in analogy with $\mathscr{H}(M) = \operatorname{colim}_k H(M \times D^k)$. A homotopical description of $\operatorname{colim}_k \Omega^k Z_k(M)$ in terms of an obscure cousin of algebraic K-theory is sought, analogous to Waldhausen's description of $\mathscr{H}(M)$ in terms of algebraic K-theory. The question comes from [22,23] but it is not explicitly formulated there.

References

1. Adams, J.F.: Infinite Loop Spaces, Annals of Mathematics Studies, vol. 90. Princeton University Press, Princeton (1978)
2. Cohen, M.: A Course in Simple-Homotopy Theory. Graduate Texts in Mathematics. Springer, Berlin (1973)
3. Cohen, M.: Simplicial structures and transverse cellularity. Ann. Math. **85**, 218–245 (1967)
4. Hatcher, A.: Higher simple homotopy theory. Ann. Math. **102**, 101–137 (1975)
5. Hatcher, A.: Concordance spaces, higher simple homotopy theory, and applications. In: Proceedings of Symposia in Pure Mathematics, vol. 32 part I, pp. 3–21 (1978)
6. Hudson, J.F.P.: Piecewise Linear Topology. W.A. Benjamin Inc., New York (1968)
7. Igusa, K.: The stability theorem for smooth pseudoisotopies. K-theory **2**, 1–355 (1988)
8. Kirby, R.C., Siebenmann, L.C.: Foundational Essays on Topological Manifolds, Smoothings, and Triangulations, with notes by John Milnor and Michael Atiyah. Annals of Mathematics Studies, vol. 88. Princeton University Press, Princeton (1977)
9. Lacher, R.C.: Cell-like spaces. Proc. Am. Math. Soc. **20**, 598–602 (1969)
10. Lacher, R.C.: Cell-like mapppings. I. Pacific J. Math. **30**, 717–731 (1969)
11. Lacher, R.C.: Cell-like mappings. II. Pacific J. Math. **35**, 649–660 (1970)
12. Mazur, B.: Relative neighborhoods and the theorems of Smale. Ann. Math. **77**, 232–249 (1963)
13. Milnor, J.: Lectures on the h-Cobordism Theorem, notes by L. Siebenmann and J. Sondow, Princeton University Press, Princeton (1965)
14. Milnor, J.: Whitehead torsion. Bull. Am. Math. Soc. **72**, 358–426 (1966)
15. Segal, G.: Classifying spaces and spectral sequences. Inst. Hautes Et. Sci. Publ. Math. **34**, 105–112 (1968)
16. Smale, S.: On the structure of manifolds. Am. J. Math. **84**, 387–399 (1962)
17. Stallings, J.R.: Lectures on Polyhedral Topology, notes by G. Ananda Swarup, Tata Institute of Fundamental Research Lectures on Mathematics, vol. 43. Tata Institute of Fundamental Research, Bombay (1967)
18. Waldhausen, F.: Algebraic K-theory of spaces. In: Algebraic and Geometric Topology (New Brunswick, NJ, 1983). Lecture Notes in Mathematics, vol. 1126, pp. 318–419. Springer, Berlin (1985)
19. Waldhausen, F.: Algebraic K-theory of spaces, a manifold approach. In: Canadian Mathematical Society Conference Proceedings, vol. 2 Part I, pp. 141–186 (1982)
20. Waldhausen, F.: Algebraic K-theory of spaces, concordance, and stable homotopy theory. In: Algebraic Topology and Algebraic K-Theory (Princeton, NJ, 1983). Annals of Mathematics Studies, vol. 113, pp. 392–417. Princeton University Press, Princeton (1987)
21. Waldhausen, F.: An outline of how manifolds relate to algebraic K-theory. In: Homotopy Theory (Durham, 1985). London Mathematical Society Lecture Note Series, vol. 117, pp. 239–247. Cambridge University Press, Cambridge (1987)
22. Weiss, M., Williams, B.: Automorphisms of manifolds and algebraic K-theory: I. K-theory **1**, 575–626 (1988)
23. Weiss, M.: Orthogonal calculus. Trans. Am. Math. Soc. **347**, 3743–3796 (1995), Erratum. Trans. Amer. Math. Soc. **350**, 851–855 (1998)
24. Whitehead, J.H.C.: Simple homotopy types. Am. J. Math. **72**, 1–57 (1952)
25. Whitehead, J.H.C.: Simplicial spaces, nuclei, and m-groups. Proc. Lond. Math. Soc. **45**, 243–327 (1939)

Jahresber Dtsch Math-Ver (2014) 115:127–128
DOI 10.1365/s13291-013-0077-z

PREFACE

Preface Issue 3/4-2013

Hans-Christoph Grunau

The year 2013 is of importance for stochastics in several respects. First of all it is the 250th anniversary of what today is known as Bayes' theorem. "An essay towards solving a problem in the doctrine of chances" by the late Rev. Mr. Bayes appeared in 1763. Thomas Bruss reviews again the context, the contents, the enormous implications and in parts also the historical discussion of this ground breaking essay. His review originates from the recently established collaboration of Zentralblatt für Mathematik and Jahresbericht der DMV.

2013 is also "The International Year of Statistics" to which the Jahresbericht contributes with the help of Winfried Stute and his survey on the statistics of point processes and their relations to risk analysis and survival analysis. After having explained the basic notions and ideas by means of simple examples from everday experiences, Winfried Stute focuses in particular on the role of point processes in market research, on modelling of purchase patterns and the influence of advertising and how statistics may help to deal with the a priori unkown model parameters.

For about 100 years general relativity has been a strong inspiration both for mathematicians and mathematics. Lorentzian geometry has been established as an active and fast developing field of mathematics which interacts not only with physics but also with other fields of mathematics like e.g. Riemannian geometry. Olaf Müller and Miguel Sánchez give a comprehensible introduction to the key features and an accessible survey on some of the most important and recent developments in Lorentzian geometry. Although there are some common features of Riemannian and Lorentzian geometry, for example the existence and uniqueness of a torsion free covariant derivative, there are however fundamental differences already at a basic level, for example

H.-Ch. Grunau (✉)
Institut für Analysis und Numerik, Fakultät für Mathematik, Otto-von-Guericke-Universität,
Postfach 4120, 39016 Magdeburg, Germany
e-mail: hans-christoph.grunau@ovgu.de

when thinking of manifolds as metric spaces, and considering notions and properties of geodesics and completeness. The authors consider, among other topics, causality, global hyperbolicity, the constraint equations for initial values of the Einstein field equations, constant mean curvature space-like hypersurfaces, singularities, definitions of mass, spinors and holonomy. Their article ends with explaining a number of conjectures and open problems. The authors see a large potential for applying Lorentzian techniques in Riemannian geometry. They mention the solution of the Yamabe problem as a prominent example for the fruitful interaction between general relativity and Riemannian geometry and advertise that still much more could be achieved in this direction.

Ulrich Felgner's historical-philosophical article reflects Hilbert's "Grundlagen der Geometrie" and its role in the long discussion, which already started in ancient Greece, on how to formulate a safe foundation of geometry.

Four extensive book reviews can be found in this double issue. Two of them are concerned with completely different aspects of Fourier Analysis. The third of the books under review addresses the interplay of symplectic and complex geometry while the fourth one deals with some recent developments in algebraic topology.

Readers and subscribers of the Jahresbericht may have noticed that this year the first two issues appeared with some delay. The last two issues have to be combined into the present single one. I apologise for all these inconveniences. However, I hope that every reader will be satisfied with the contents of the articles and reviews and will enjoy this issue.