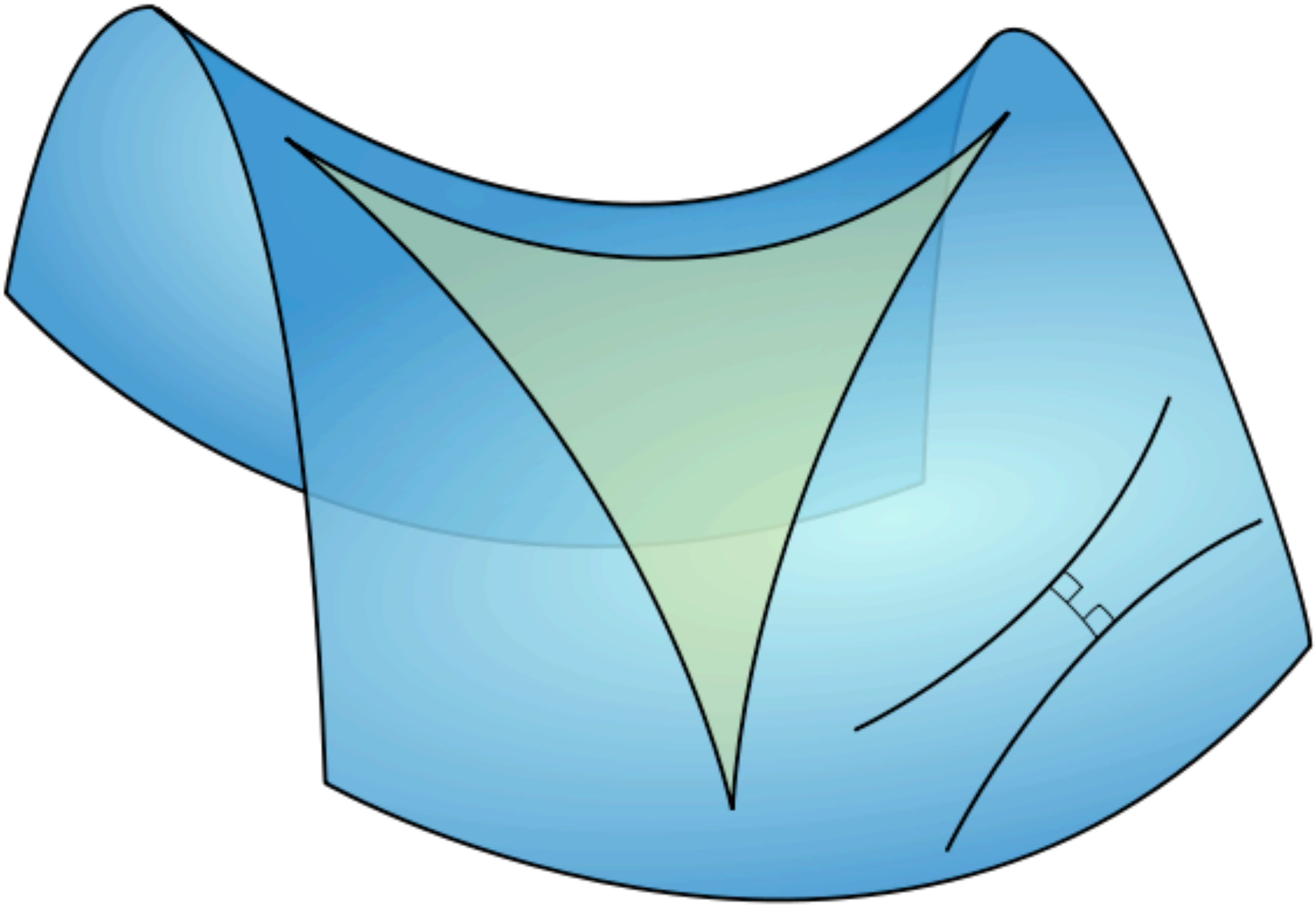
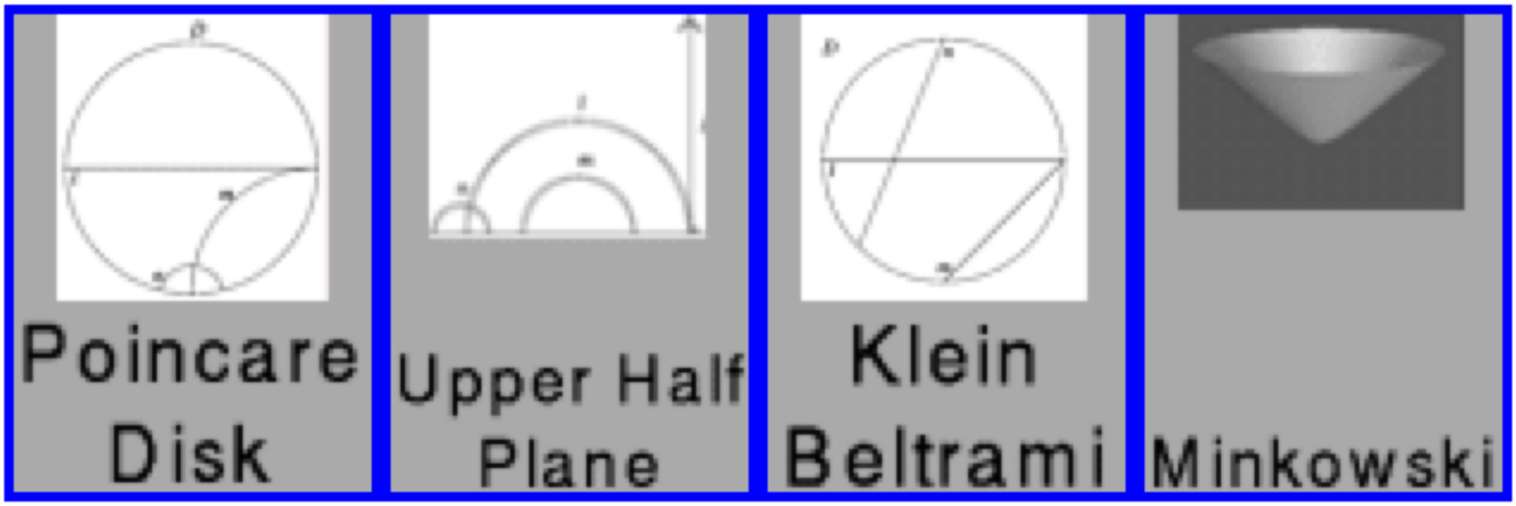
How can we project the hyperbolic plane onto the Euclidian plane?



Four different projections are used



Hyperbolic Geometry Websites

http://www.maths.gla.ac.uk/~wws/cabripages/hyperbolic/experiment.html

http://www.geom.uiuc.edu/~crobles/hyperbolic/hypr/modl/

http://www.geom.uiuc.edu/~crobles/hyperbolic/hypr/ibm/

http://cs.unm.edu/~joel/NonEuclid/NonEuclid.html

Hyperbolic Tesselations

http://aleph0.clarku.edu/~djoyce/poincare/poincare.html

http://aleph0.clarku.edu/~djoyce/poincare/other_tess.html

http://www.plunk.org/~hatch/HyperbolicTesselations

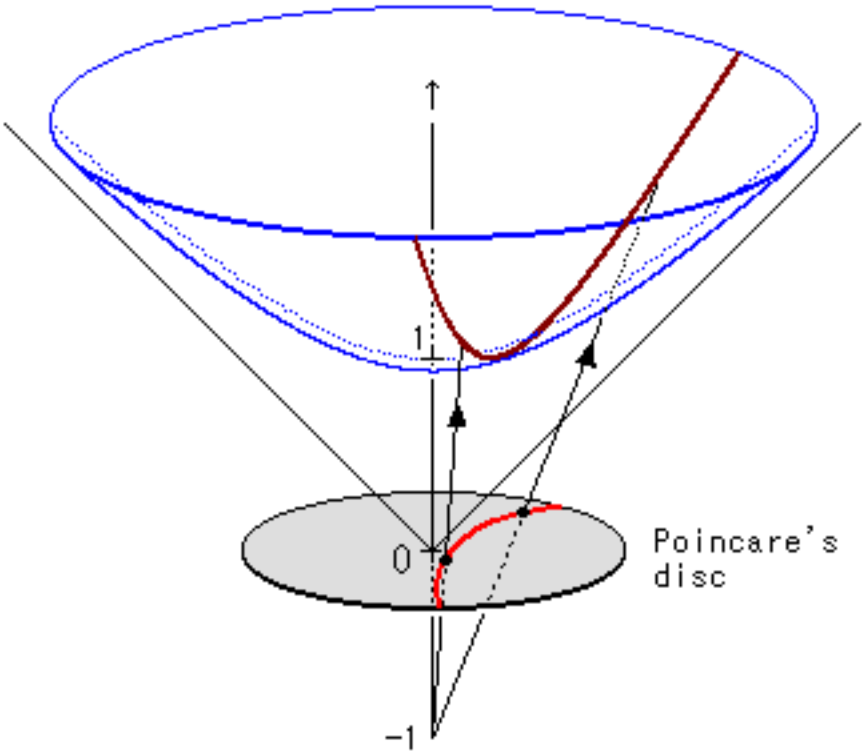
Hyperbolic Computer Graphics

http://graphics.stanford.edu/papers/h3cga/

http://www.geom.uiuc.edu/docs/research/webviz/webviz/node2.html

http://sigchi.org/chi95/proceedings/papers/jl_bdy.htm

[2]



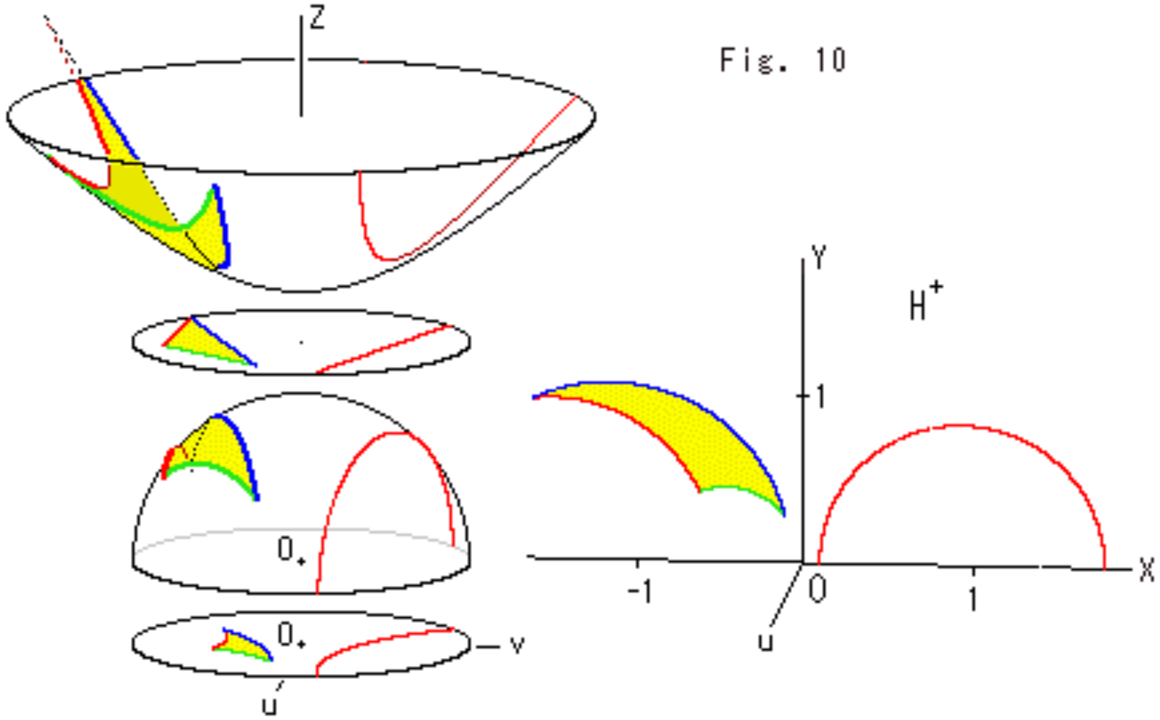
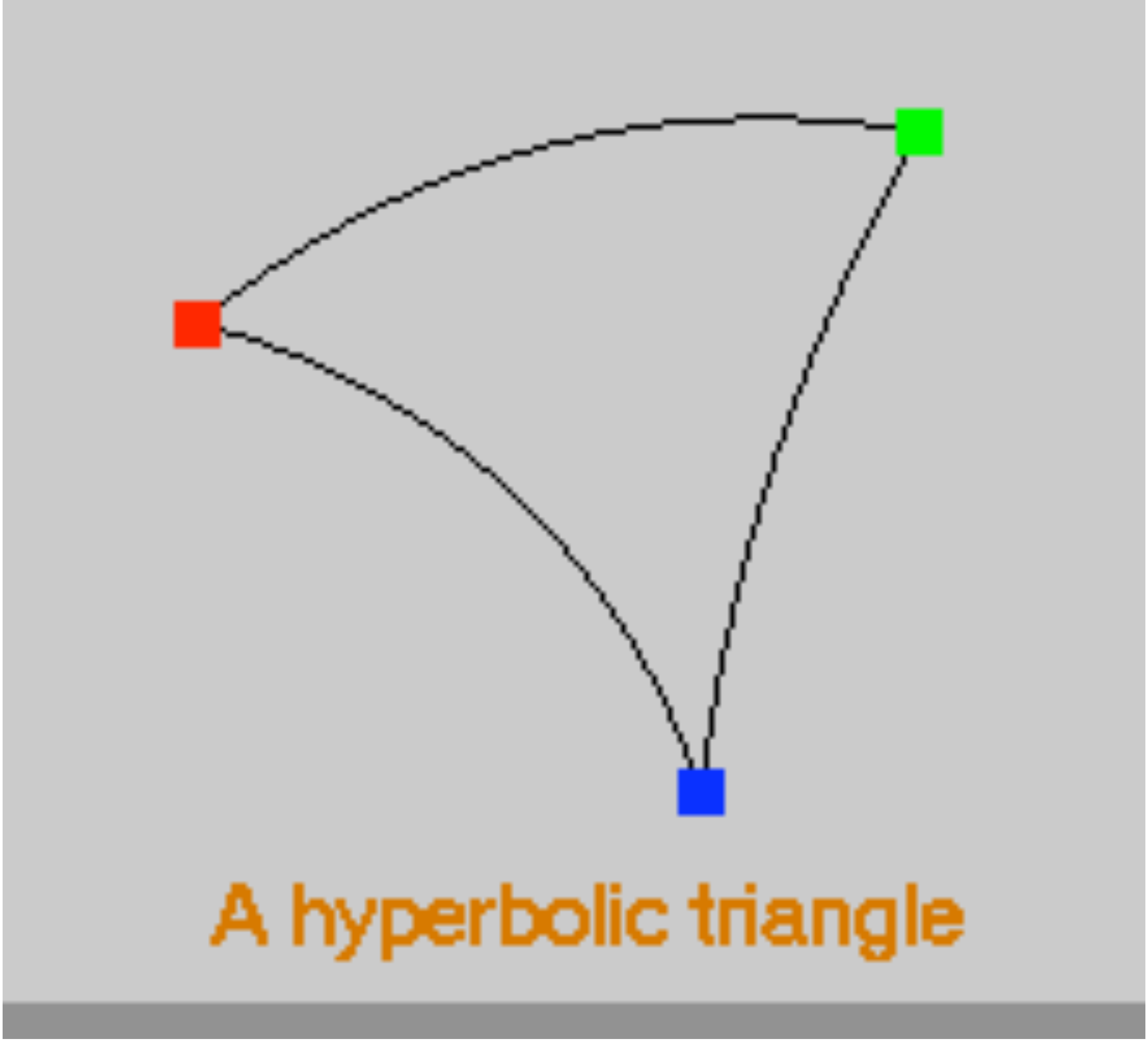
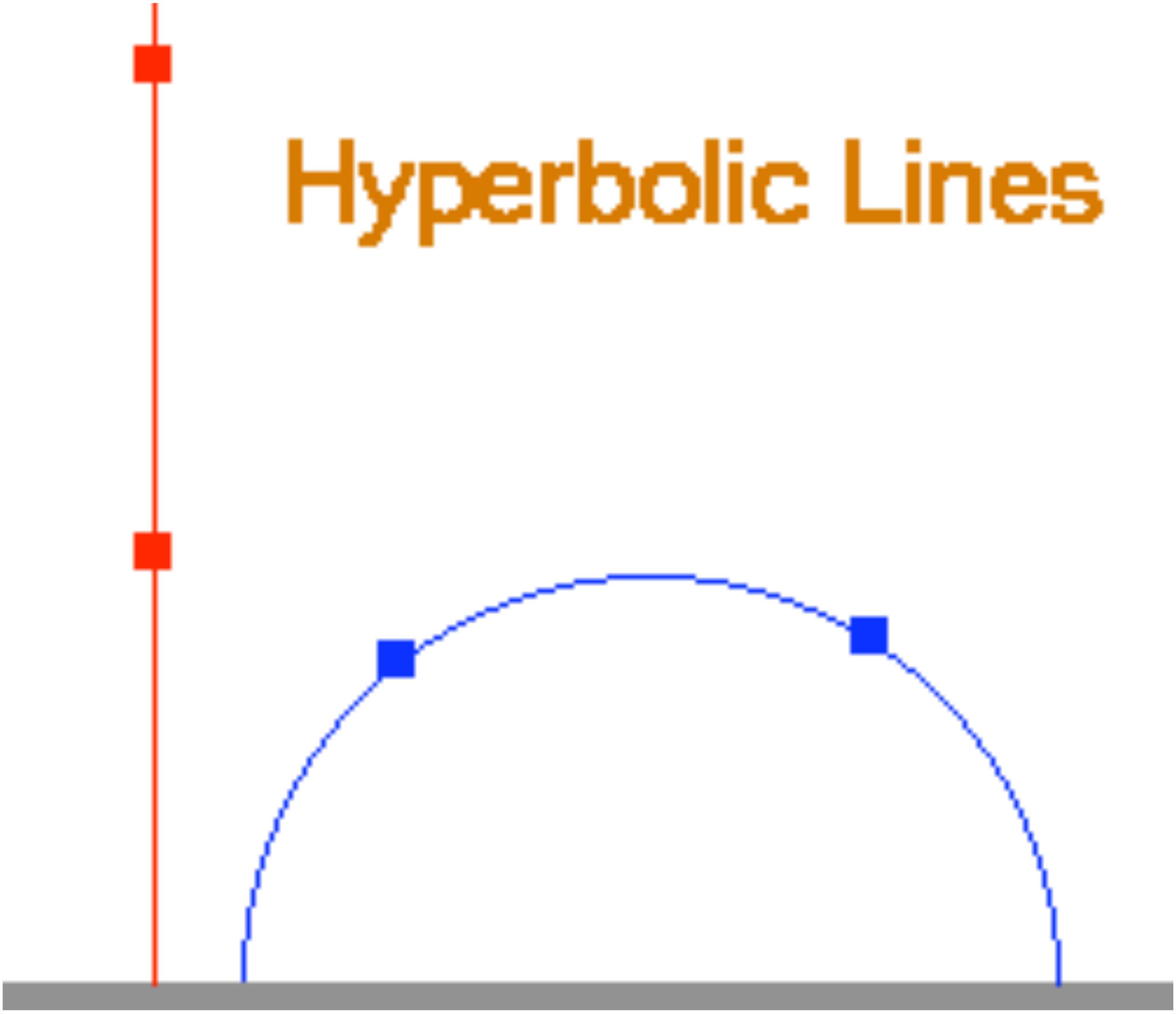


Fig. 10


A hyperbolic triangle



Hyperbolic Lines

Cinderella Story

Hyperbolic geometry is the Cinderella story of mathematics. Rejected and hidden while her two sisters (spherical and euclidean geometry) hogged the limelight, hyperbolic geometry was eventually rescued and emerged to outshine them both. The first part of this saga---how Bolyai and Lobachevsky laboured in vain to win recognition for their subject---is well known, and English translations of the key documents are available in Bonola's classic Non-Euclidean Geometry. However, the turning point of the story has not been documented in English until now.

Beltrami came to the rescue of hyperbolic geometry in 1868 by interpreting it on a surface of constant negative curvature. By giving a concrete meaning to the hyperbolic plane, he put Bolyai's and Lobachevsky's work on a sound logical foundation for the first time, and showed that it was a part of classical differential geometry. This was quickly followed by interpretations in projective geometry by Klein in 1871, and in the complex numbers by Poincare in 1882.

Hyperbolic geometry had arrived, and with Poincare it joined the main-stream of mathematics. He used it immediately in differential equations, complex analysis, and number theory, and its place has been secure in these disciplines ever since. He also began to use it in low-dimensional topology, an idea kept alive by a handful of topologists until the spectacular blossoming of this field under Thurston in the late 1970s. Now, hyperbolic geometry is the generic geometry in dimensions 2 and 3.

"For God's sake, please give it up. Fear it no less than the sensual passions, because it, too, may take up all your time and deprive you of your health, peace of mind and happiness in life."

Wolfgang Bolyai to his son Janos Bolyai
on the dangers of hyperbolic geometry

What exactly is a hyperbolic plane?

DH: There are many ways of describing the hyperbolic plane. In formal geometric terms, it is "a simply connected Riemannian manifold with negative Gaussian curvature." In higher-level mathematics courses it is often defined as the geometry that is described by the upper half-plane model. One way of understanding it is that it's the geometric opposite of the sphere. On a sphere, the surface curves in on itself and is closed. A hyperbolic plane is a surface in which the space curves away from itself at every point. Like a Euclidean plane it is open and infinite, but it has a more complex and counterintuitive geometry.

The hyperbolic plane is sometimes described as a surface in which the space expands?

DH: Actually that is true for many spaces, but it's true for hyperbolic space in a particular way. Consider how circles on a surface behave. If you think of a series of circles around a point on a regular Euclidean plane, as you draw larger circles the length of the circumference increases linearly. Now on a hyperbolic plane, the circumference of the circles doesn't just increase linearly, but exponentially. The perimeter and also the area of the circles gets bigger much faster. On a sphere, the circles get larger at first, but then as you go further they actually begin to get smaller. On a sphere the circumference of a circle is always less than 2πr, on a hyperbolic plane it is more. A similar thing happens with the area, which also increases much faster in hyperbolic space.

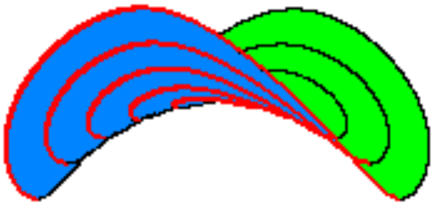
Spherical geometry is pretty easy to understand because we see spheres all around us. But when mathematicians first started to study hyperbolic geometry, they didn't have any idea what this space might look like and they were nearly driven mad trying to understand this space. In 1997, Daina, you worked out how to make a physical model of the hyperbolic plane using crochet. How did that discovery come about?

DT: For the past 125 years or so mathematicians had conceptual models of the hyperbolic plane, such as the Poincaré disc model, developed by the French mathematician Henri Poincaré. Some of the models had great aesthetic appeal, especially given the enormous variety of repeating patterns that are possible in the hyperbolic plane. After the geometer Donald Coxeter explained these conceptual models to Escher, he used patterns based on these models in several of his prints.

Does that mean geometry is going to have some sort of renaissance?

DH: Oh yes, geometry is having a renaissance. One indication of this is the number of geometry courses being taught around the country. For example, when I first came to Cornell there was only one undergraduate geometry course; now we have eight. A further indication is the work of William Thurston, who has hypothesized a classification for the different types of three-dimensional manifolds - the three-dimensional analogues of two-dimensional surfaces (such as the surface of a sphere or a donut). Where two-dimensional geometry comes in just three types---the Euclidean plane, the sphere, and the hyperbolic plane---Thurston's Hypothesis says that there are eight distinct 3-D spatial types. Over the past year the mathematical world has been excited by the news that Thurston's Hypothesis might have been proved. If that turns out to be true, then mathematics most famous geometric problem - the Poincaré Conjecture - will also have been proved, and there's a million dollar prize attached to that discovery.

Fig. 3



Look at Fig. 3.

There is a term **curvature**
is zero. A hyperbolic plane
geometry, Euclidean geom
The hyperbolic plane is, of
way. Fig. 3 shows this con
make to treat a plane as a plane? This i

Fig. 4

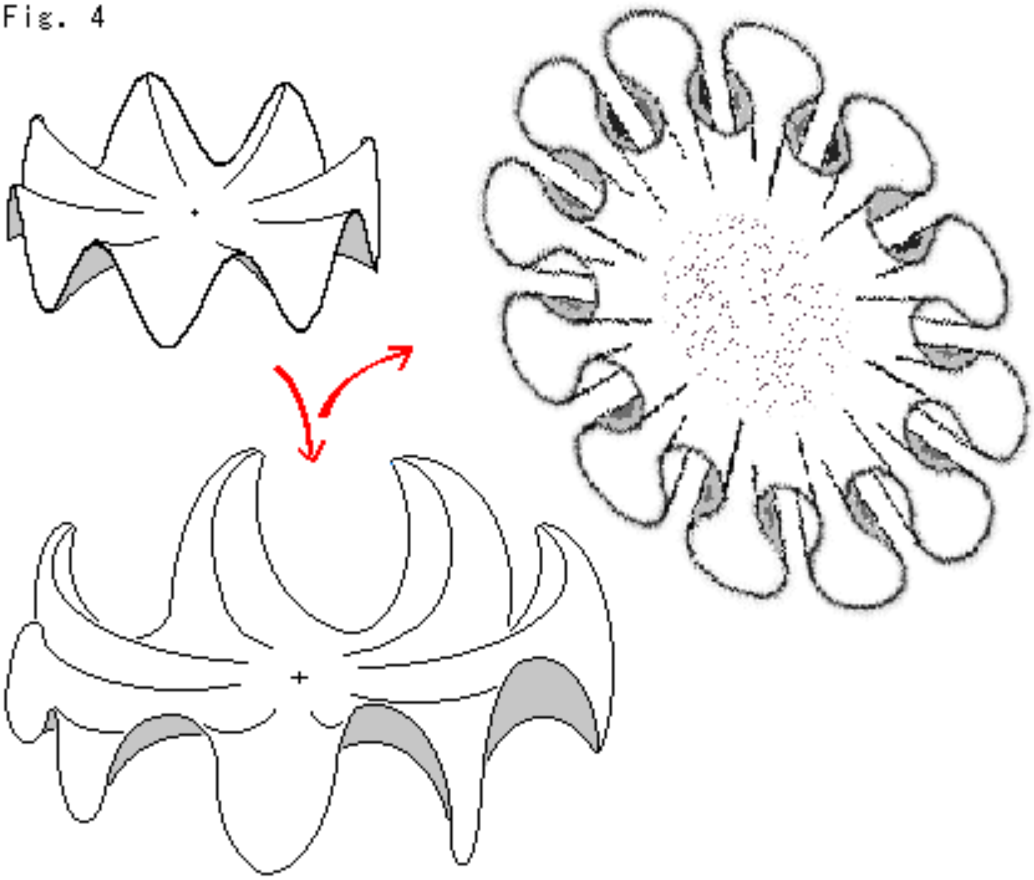


Fig. 5

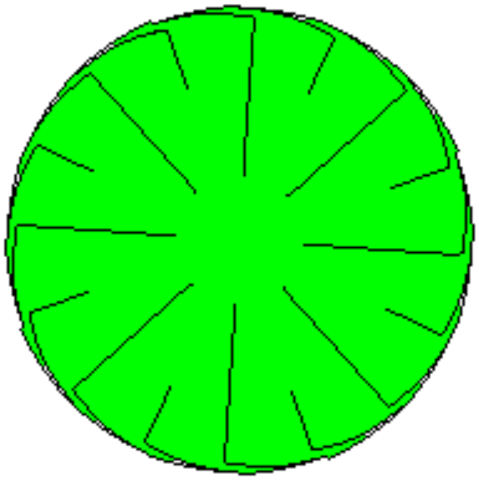
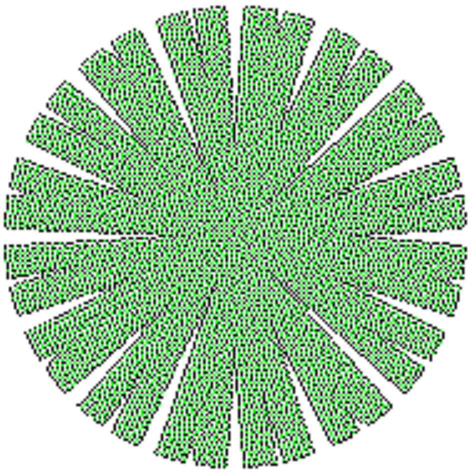


Fig. 5’



Crocheting What Euclid Couldn't Grasp

A visiting scholar at Cornell University has taken both the math world and, more recently, the art world by storm with a touchable form of advanced geometry. Daina Taimina, a mathematician and crocheter, discovered a way to create durable and easy-to-use models of hyperbolic space. People have been attempting this ever since the concept emerged in the early 1800s and overturned Euclidean geometry's assumptions about parallel lines.

Students are excited by the crochet creations, some of which resemble curvy leaves of lettuce or kale. Straight lines of one color of yarn are stitched into models made of another color. By bending the forms, students see a new three-dimensional relationship between parallel lines.

"Maybe there's something psychological about using the yarns and the handicraft.... It just takes out the anxiety. It's not the same as looking at some abstract formula," Dr. Taimina says. "People say, 'It's very important that we can touch it - we can learn through feeling.' "

Hilbert proved that it is impossible to represent the hyperbolic plane in three-dimensional space analytically.

Hyperbolic geometry, conceived by mathematician Carl Gauss in 1816, is stranger still. Like planar geometry, it posits that the shortest distance between two points is a straight line. And hyperbolic space, like spherical space, has a constant curvature—except the curvature is negative rather than positive. Hyperbolic geometry describes a world that is curving away from itself at every point, making it the precise opposite of a sphere, whatever that might look like. (One is tempted to picture an inside-out sphere, but that still describes a positive curvature, since space is curving toward itself at each point.)

The rediscovery of hyperbolic space was not greeted enthusiastically by the analytically oriented German and Austrian mathematicians who dominated mathematics in the West; they dreamed of a logical, orderly universe that could be represented through equations. Not until very recently—after the fall of the iron curtain—did the strange and illogical beauty of hyperbolic forms emerge yet again to claim the attention of mathematicians.

I ask Henderson how it is that shapes that cannot be imagined nonetheless can be found in his wife's knitting bowl. "A hundred years ago, the mathematician David Hilbert proved a theorem that it is impossible to represent the hyperbolic plane in three-dimensional space analytically," he says. " 'Analytically' means 'with equations.' Everybody left off the word analytically later on. They were worried that mistakes or errors would creep into mathematics through geometric intuition, and so they discouraged the study of geometry and everything associated with this weird kind of thinking."

http://theiff.org/oexhibits/oe1.html
http://discovermagazine.com/2006/mar/knit-theory
http://www.news.cornell.edu/stories/Dec05/Taimina.hyperbolic.lg.html

Heisenberg’s first graduate student was Felix Bloch. One day, while walking together, they started to discuss the concepts of space and time. Bloch had just finished reading Weyl’s book Space, Time, Matter, the same book that Heisenberg had read as a young man. Still very much under the influence of this scholarly work, Bloch declared that he now understood that space was simply the field of affine transformations.

Heisenberg paused, looked at him, and replied:

“Nonsense, space is blue and birds fly through it.”

After Poincare

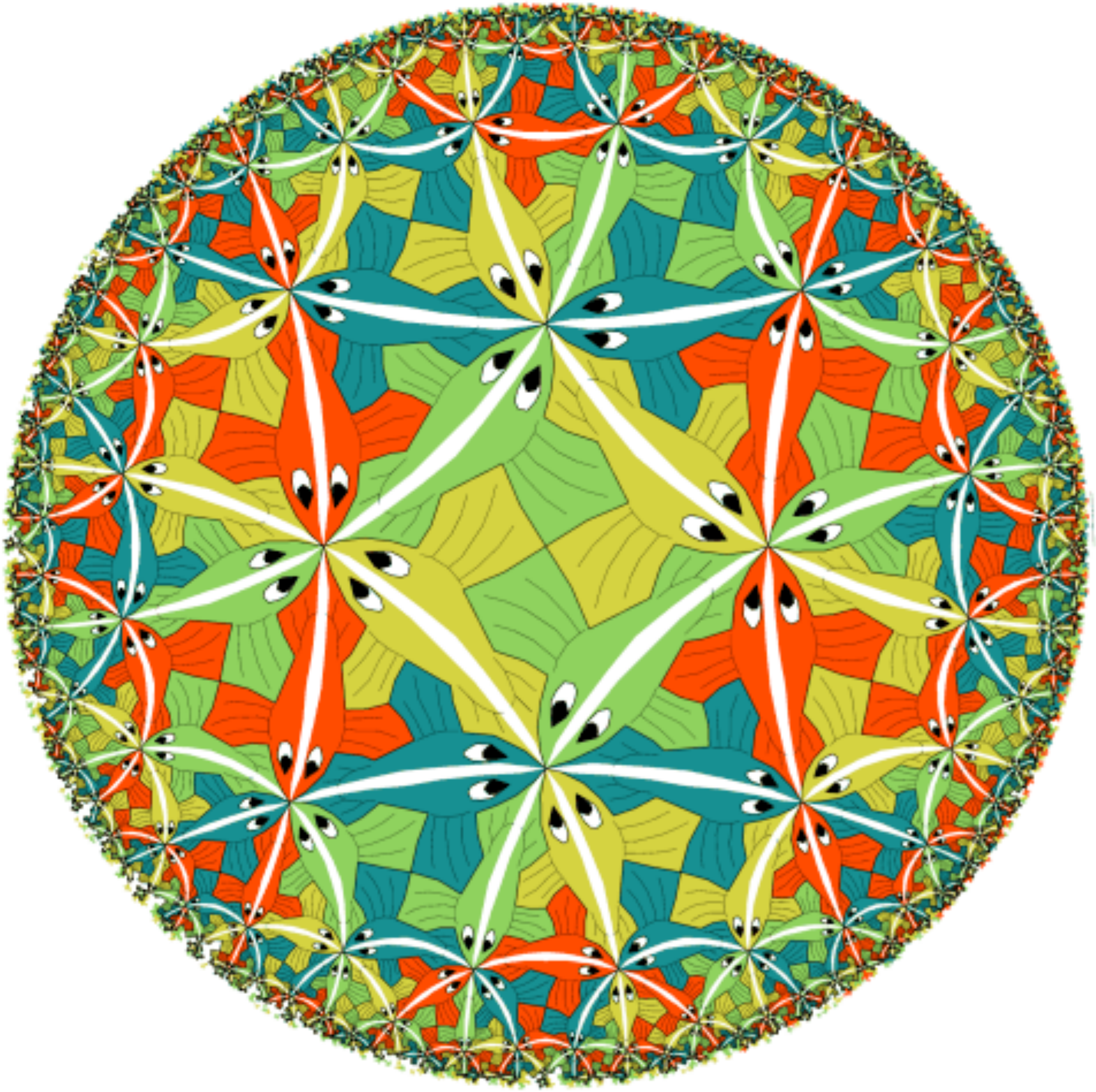
Hyperbolic geometry took a new turn with the advent of Poincare. Beltrami and Klein were primarily geometers, who used known geometry to construct realisations of what was, until then, an unknown geometry. Poincare began in other fields, and made the surprising discovery that hyperbolic geometry was already present in mainstream mathematics. In a famous passage, he described how the revelation came to him:

Just at this time I left Caen, where I was then living, to go on a geological excursion under the auspices of the school of mines. The changes of travel made me forget about my mathematical work. Having reached Coutances, we entered an omnibus to go some place or other. At the moment when I put my foot on the step the idea came to me, without anything in my former thoughts seeming to have paved the way for it, that the transformations I had used to define Fuchsian functions were identical with those of non-Euclidean geometry. I did not verify the idea: I should not have had time, as, upon taking my seat in the omnibus, I went on with a conversation already commenced, but I felt a perfect certainty. On my return to Caen, for conscience' sake I verified the result at my leisure.

Then I turned my attention to the study of some arithmetic questions apparently without success and without a suspicion of any connection with my previous researches. Disgusted with my failure, I went to spend a few days at the seaside, and thought of something else. One morning, walking along the bluff, the idea came to me, with just the same characteristics of brevity, suddenness and immediate certainty, that the arithmetic tranformations of ternary quadratic forms were identical with those of non-Euclidean geometry.

This discovery cast hyperbolic geometry in an entirely new light. Now it could be seen as not merely logically valid, but also natural, familiar and potentially useful. Poincare proceeded to show how useful it was with a slew of applications to complex analysis, differential equations, number theory and later topology.

Poincare developed new models of hyperbolic geometry: two long papers on the group theory of differential equations and a third, which is not so well known on the applications of hyperbolic geometry to number theory. As we shall see, number theory contains possibly the earliest hints of hyperbolic geometry in mainstream mathematics, some being visible in the memoir of Lagrange on quadratic forms.



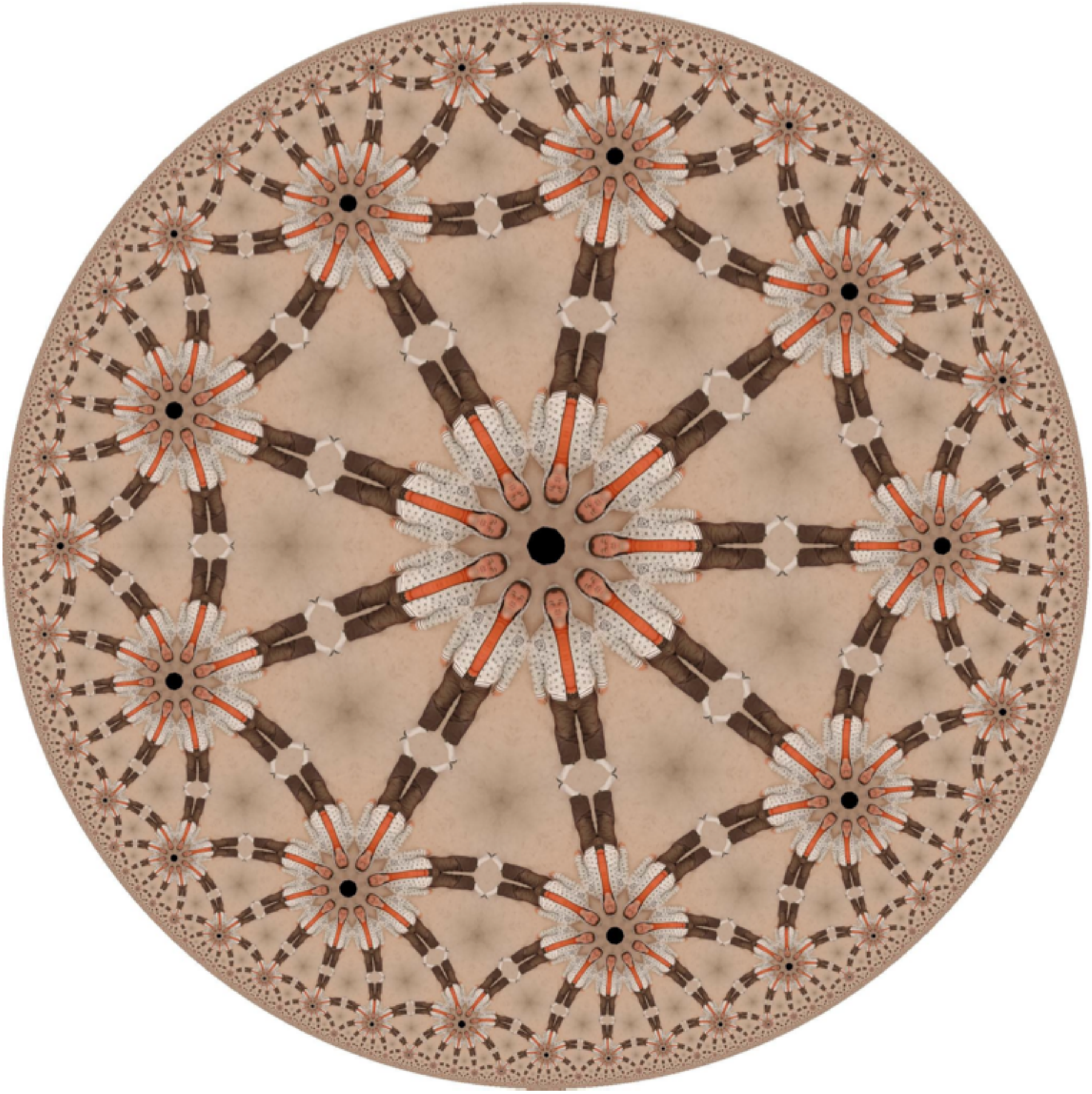


FIG. 1: The exponentially growing number of people lying on the hyperbolic floor illustrates the exponential expansion of hyperbolic space. All people are of the same hyperbolic size. The `Poincaré` tool developed by Bill Horn is used to construct the tessellation of the hyperbolic plane in the Poincaré disc model with the Schläfli symbol $\{9, 3\}$, rendering an image of the last author.

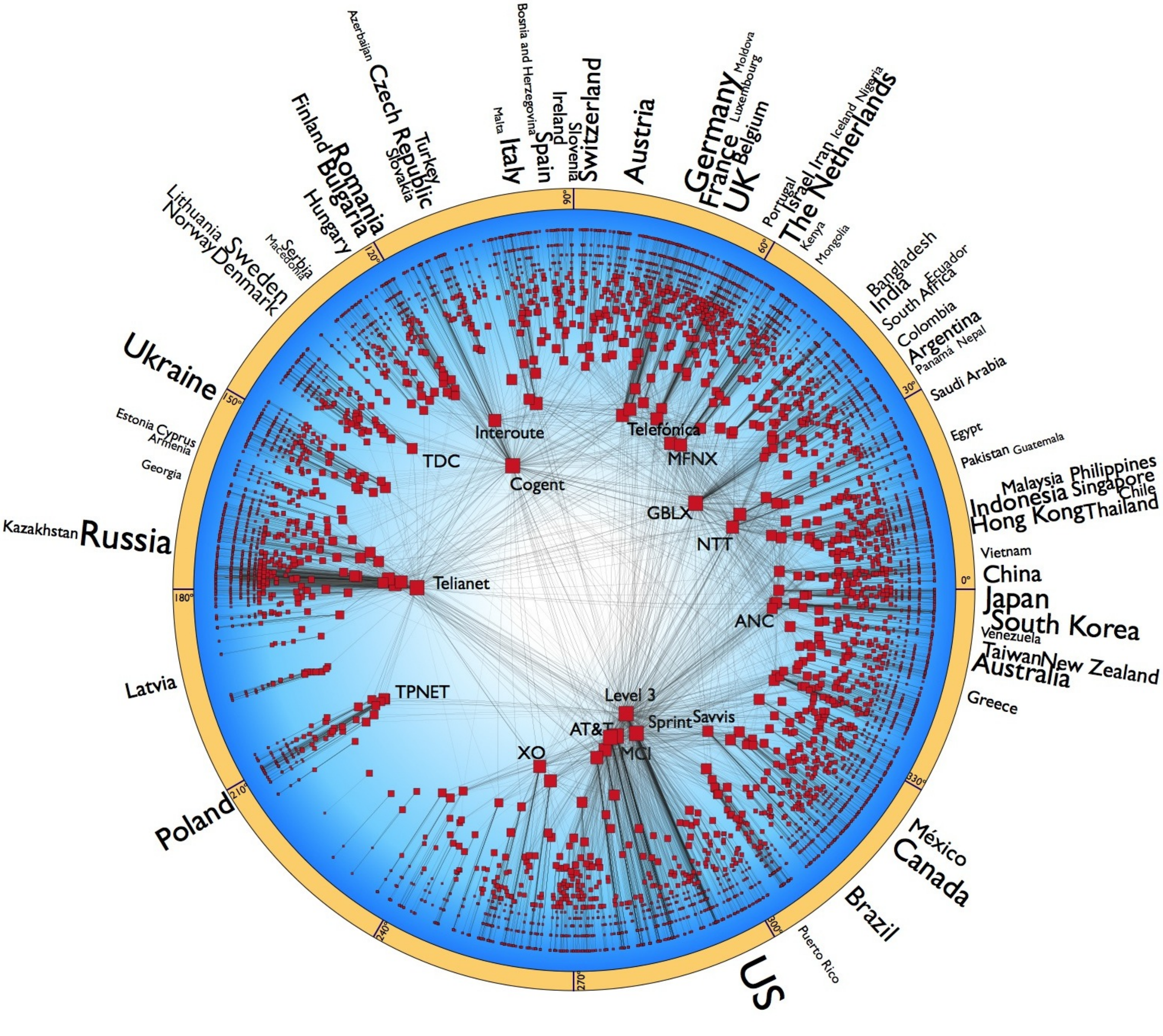
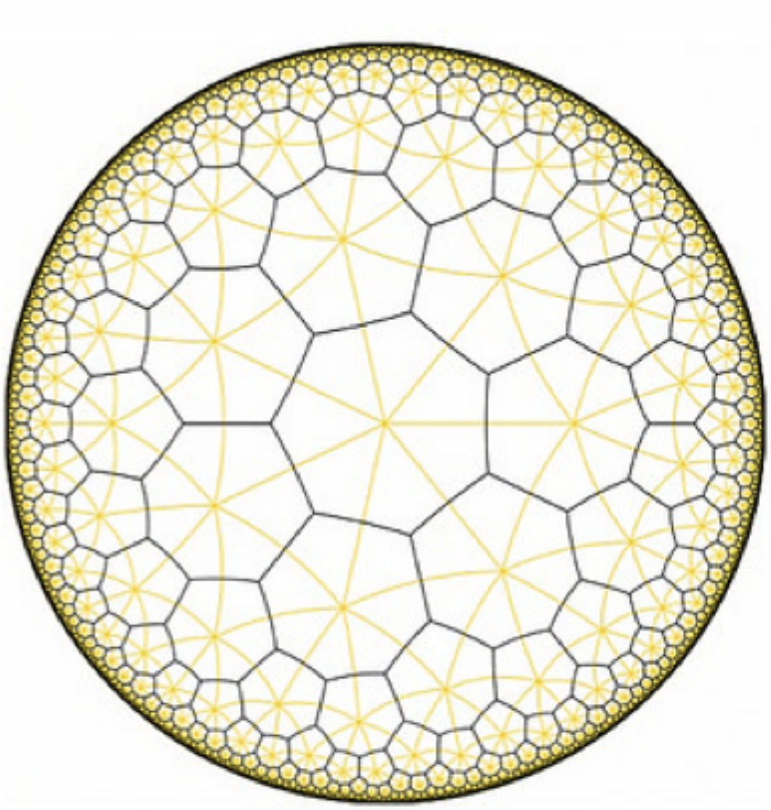
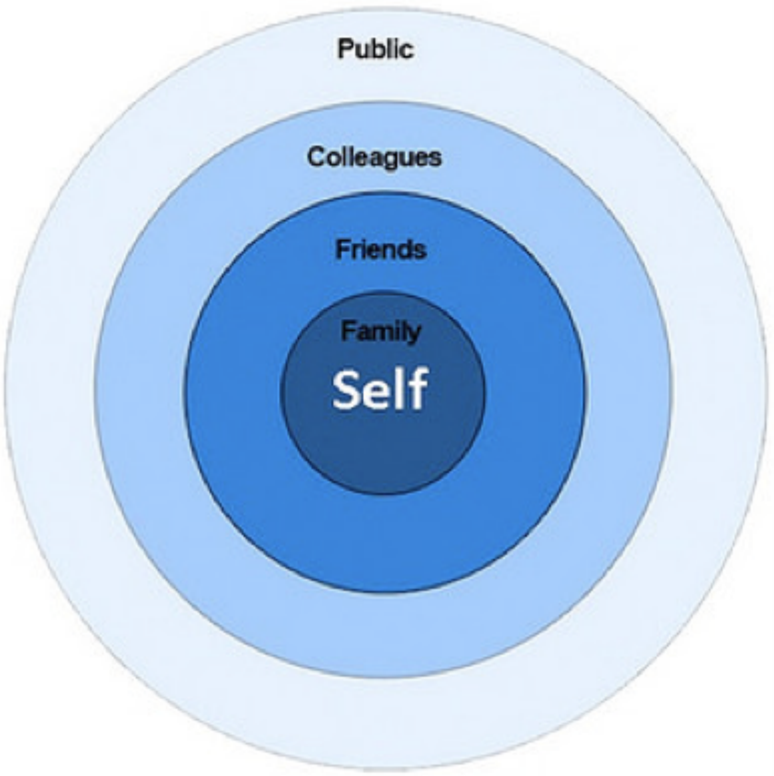


FIG. 3: The hyperbolic map of the Internet is similar to a synthetic Einsteinian network in Fig. 2. The size of AS nodes is proportional to the logarithm of their degrees. For the sake of clarity, only ASs with degree above 3, and only the connections with probability $p(x) > 0.5$ given by Eq. (2) are shown. The font size of the country names is proportional to the logarithm of the number of ASs that the country has. Only the names of countries with more than 10 ASs are included. The methods used to map ASs to their countries are described in Appendix D.

single connected component. Using this property along with the statistical independence of the graph edges, it becomes possible to infer coordinates of ASs in $\mathcal{G}(k_T)$ ignoring the remainder of the AS graph. This property is practically important because the size of $\mathcal{G}(k_T)$ decreases very fast as k_T increases, which speeds up likelihood maximisation algorithms tremendously. In a nutshell, our method starts with a subgraph $\mathcal{G}(k_T)$ small enough for standard maximisation algorithms being able to reliably and quickly infer the coordinates of ASs in $\mathcal{G}(k_T)$. Once these are found, we gradually increase k_T to iteratively add layers of lower-degree ASs. While doing so, we use the already inferred AS coordinates as a reference frame to assign initial coordinates to newly added ASs. This initial coordinate assignment significantly improves the convergence time of maximisation algorithms. All other details of our mapping method can be found in Appendix B.

Social Networks





The Poincare Conjecture

With no mass, we are left with "pure" space. This is a 3-D manifold that has some sort of geometry. With our now homogeneous 3-D manifold (the inhomogeneous curving effects of mass have been mitigated), the possible geometries are analogous to the geometries of the surfaces of the 2-D objects that we examined before. Now, however, instead of being polygons, the tiles will be polyhedra, and the type or types of polyhedra that will tile or "pack" a particular space are determined by the specific geometry of that space.

So, the question remains: what are the possible geometries? In 1982 William Thurston, one of the most influential modern geometers and topologists, proposed that there are eight possible geometries, Euclidean, spherical, hyperbolic, and five other systems. In the early and mid-2000s Perelman proved Thurston's claims to be correct. This result also proved the Poincaré conjecture, which considered only spherical 3-D manifolds.

Now that Thurston's geometrization conjecture has been proven to be correct— and has earned the title of "theorem"—we have essentially a complete list of possibilities for the fundamental geometry of our space. The task now is to determine which geometry actually governs the real world we live in. This is essentially what Gauss tried to do on his mountaintops so many years ago. The problem, as we have seen, is that to reach a definitive answer, we need to be able to look at extremely large shapes, much larger than anything on Earth or even in our galaxy, perhaps.

So we are, indeed, much like the ant on its surface: we know what is happening with the local curvature, but we are looking too closely to be able to discern much about the large-scale geometry of our system. If we had to guess the specific geometry of our space, we, like the ant, would do well to guess hyperbolic. Indeed, Thurston's Geometrization Theorem confirms that most spaces are spaces that obey the "many parallels" version of Euclid's fifth postulate.

http://www.youtube.com/watch?v=qsE2UKkIKXU

The Poincare Conjecture and the Thurston Conjecture become the Poincare Theorem and the Thurston Theorem

http://en.wikipedia.org/wiki/Poincare_conjecture

http://en.wikipedia.org/wiki/Solution_of_the_Poincare_conjecture

http://www.claymath.org/poincare/

http://comet.lehman.cuny.edu/sormani/others/perelman/introperelman.html

http://www.math.unl.edu/~mbrittenham2/ldt/poincare.html

The New York Times
nytimes.com

August 15, 2006

Elusive Proof, Elusive Prover: A New Mathematical Mystery

By DENNIS OVERBYE

Correction Appended

Grisha Perelman, where are you?

Three years ago, a Russian mathematician by the name of Grigory Perelman, a k a Grisha, in St. Petersburg, announced that he had solved a famous and intractable mathematical problem, known as the Poincaré conjecture, about the nature of space.

After posting a few short papers on the Internet and making a whirlwind lecture tour of the United States, Dr. Perelman disappeared back into the Russian woods in the spring of 2003, leaving the world's mathematicians to pick up the pieces and decide if he was right.

Now they say they have finished his work, and the evidence is circulating among scholars in the form of three book-length papers with about 1,000 pages of dense mathematics and prose between them.

As a result there is a growing feeling, a cautious optimism that they have finally achieved a landmark not just of mathematics, but of human thought.

"It's really a great moment in mathematics," said Bruce Kleiner of Yale, who has spent the last three years helping to explicate Dr. Perelman's work. "It could have happened 100 years from now, or never."

In a speech at a conference in Beijing this summer, Shing-Tung Yau of Harvard said the understanding of three-dimensional space brought about by Poincaré's conjecture could be one of the major pillars of math in the 21st century.

Quoting Poincaré himself, Dr.Yau said, "Thought is only a flash in the middle of a long night, but the flash that means everything."

But at the moment of his putative triumph, Dr. Perelman is nowhere in sight. He is an odds-on favorite to win a Fields Medal, math's version of the Nobel Prize, when the International Mathematics Union convenes in Madrid next Tuesday. But there is no indication whether he will show up.

Also left hanging, for now, is $1 million offered by the Clay Mathematics Institute in Cambridge, Mass., for the first published proof of the conjecture, one of seven outstanding questions for which they offered a ransom back at the beginning of the millennium.

"It's very unusual in math that somebody announces a result this big and leaves it hanging," said John Morgan of Columbia, one of the scholars who has also been filling in the details of Dr. Perelman's work.

Mathematicians have been waiting for this result for more than 100 years, ever since the French polymath Henri Poincaré posed the problem in 1904. And they acknowledge that it may be another 100 years before its full implications for math and physics are understood. For now, they say, it is just beautiful, like art or a challenging new opera.

Dr. Morgan said the excitement came not from the final proof of the conjecture, which everybody felt was true, but the method, "finding deep connections between what were unrelated fields of mathematics."

William Thurston of Cornell, the author of a deeper conjecture that includes Poincaré's and that is now apparently proved, said, "Math is really about the human mind, about how people can think effectively, and why curiosity is quite a good guide," explaining that curiosity is tied in some way with intuition.

"You don't see what you're seeing until you see it," Dr. Thurston said, "but when you do see it, it lets you see many other things."

Depending on who is talking, Poincaré's conjecture can sound either daunting or deceptively simple. It asserts that if any loop in a certain kind of three-dimensional space can be shrunk to a point without ripping or tearing either the loop or the space, the space is equivalent to a sphere.

The conjecture is fundamental to topology, the branch of math that deals with shapes, sometimes described as geometry without the details. To a topologist, a sphere, a cigar and a rabbit's head are all the same because they can be deformed into one another. Likewise, a coffee mug and a doughnut are also the same because each has one hole, but they are not equivalent to a sphere.

In effect, what Poincaré suggested was that anything without holes has to be a sphere. The one qualification was that this "anything" had to be what mathematicians call compact, or closed, meaning that it has a finite extent: no matter how far you strike out in one direction or another, you can get only so far away before you start coming back, the way you can never get more than 12,500 miles from home on the Earth.

In the case of two dimensions, like the surface of a sphere or a doughnut, it is easy to see what Poincaré was talking about: imagine a rubber band stretched around an apple or a doughnut; on the apple, the rubber band can be shrunk without limit, but on the doughnut it is stopped by the hole.

With three dimensions, it is harder to discern the overall shape of something; we cannot see where the holes might be. "We can't draw pictures of 3-D spaces," Dr. Morgan said, explaining that when we envision the surface of a sphere or an apple, we are really seeing a two-dimensional object embedded in three dimensions. Indeed, astronomers are still arguing about the overall shape of the universe, wondering if its topology resembles a sphere, a bagel or something even more complicated.

Poincaré's conjecture was subsequently generalized to any number of dimensions, but in fact the three-dimensional version has turned out to be the most difficult of all cases to prove. In 1960 Stephen Smale, now at the Toyota Technological Institute at Chicago, proved that it is true in five or more dimensions and was awarded a Fields Medal. In 1983, Michael Freedman, now at Microsoft, proved that it is true in four dimensions and also won a Fields.

“You get a Fields Medal for just getting close to this conjecture,” Dr. Morgan said.

In the late 1970’s, Dr. Thurston extended Poincaré’s conjecture, showing that it was only a special case of a more powerful and general conjecture about three-dimensional geometry, namely that any space can be decomposed into a few basic shapes.

Mathematicians had known since the time of Georg Friedrich Bernhard Riemann, in the 19th century, that in two dimensions there are only three possible shapes: flat like a sheet of paper, closed like a sphere, or curved uniformly in two opposite directions like a saddle or the flare of a trumpet. Dr. Thurston suggested that eight different shapes could be used to make up any three-dimensional space.

“Thurston’s conjecture almost leads to a list,” Dr. Morgan said. “If it is true,” he added, “Poincaré’s conjecture falls out immediately.” Dr. Thurston won a Fields in 1982.

Topologists have developed an elaborate set of tools to study and dissect shapes, including imaginary cutting and pasting, which they refer to as “surgery,” but they were not getting anywhere for a long time.

In the early 1980’s Richard Hamilton of Columbia suggested a new technique, called the Ricci flow, borrowed from the kind of mathematics that underlies Einstein’s general theory of relativity and string theory, to investigate the shapes of spaces.

Dr. Hamilton’s technique makes use of the fact that for any kind of geometric space there is a formula called the metric, which determines the distance between any pair of nearby points. Applied mathematically to this metric, the Ricci flow acts like heat, flowing through the space in question, smoothing and straightening all its bumps and curves to reveal its essential shape, the way a hair dryer shrink-wraps plastic.

Dr. Hamilton succeeded in showing that certain generally round objects, like a head, would evolve into spheres under this process, but the fates of more complicated objects were problematic. As the Ricci flow progressed, kinks and neck pinches, places of infinite density known as singularities, could appear, pinch off and even shrink away. Topologists could cut them away, but there was no guarantee that new ones would not keep popping up forever.

“All sorts of things can potentially happen in the Ricci flow,” said Robert Greene, a mathematician at the University of California, Los Angeles. Nobody knew what to do with these things, so the result was a logjam.

It was Dr. Perelman who broke the logjam. He was able to show that the singularities were all friendly. They turned into spheres or tubes. Moreover, they did it in a finite time once the Ricci flow started. That meant topologists could, in their fashion, cut them off, and allow the Ricci process to continue to its end, revealing the topologically spherical essence of the space in question, and thus proving the conjectures of both Poincaré and Thurston.

Dr. Perelman’s first paper, promising “a sketch of an eclectic proof,” came as a bolt from the blue when it was posted on the Internet in November 2002. “Nobody knew he was working on the Poincaré conjecture,” said Michael T. Anderson of the State University of New York in Stony Brook.

Dr. Perelman had already established himself as a master of differential geometry, the study of curves and surfaces, which is essential to, among other things, relativity and string theory Born in St. Petersburg in 1966, he distinguished himself as a high school student by winning a gold medal with a perfect score in the International Mathematical Olympiad in 1982. After getting a Ph.D. from St. Petersburg State, he joined the Steklov Institute of Mathematics at St. Petersburg.

In a series of postdoctoral fellowships in the United States in the early 1990’s, Dr. Perelman impressed his colleagues as “a kind of unworldly person,” in the words of Dr. Greene of U.C.L.A. — friendly, but shy and not interested in material wealth.

“He looked like Rasputin, with long hair and fingernails,” Dr. Greene said.

Asked about Dr. Perelman’s pleasures, Dr. Anderson said that he talked a lot about hiking in the woods near St. Petersburg looking for mushrooms.

Dr. Perelman returned to those woods, and the Steklov Institute, in 1995, spurning offers from Stanford and Princeton, among others. In 1996 he added to his legend by turning down a prize for young mathematicians from the European Mathematics Society.

Until his papers on Poincaré started appearing, some friends thought Dr. Perelman had left mathematics. Although they were so technical and abbreviated that few mathematicians could read them, they quickly attracted interest among experts. In the spring of 2003, Dr. Perelman came back to the United States to give a series of lectures at Stony Brook and the Massachusetts Institute of Technology, and also spoke at Columbia, New York University and Princeton.

But once he was back in St. Petersburg, he did not respond to further invitations. The e-mail gradually ceased.

“He came once, he explained things, and that was it,” Dr. Anderson said. “Anything else was superfluous.”

Recently, Dr. Perelman is said to have resigned from Steklov. E-mail messages addressed to him and to the Steklov Institute went unanswered.

In his absence, others have taken the lead in trying to verify and disseminate his work. Dr. Kleiner of Yale and John Lott of the University of Michigan have assembled a monograph annotating and explicating Dr. Perelman’s proof of the two conjectures.

Dr. Morgan of Columbia and Gang Tian of Princeton have followed Dr. Perelman’s prescription to produce a more detailed 473-page step-by-step proof only of Poincaré’s Conjecture. “Perelman did all the work,” Dr. Morgan said. “This is just explaining it.”

Both works were supported by the Clay institute, which has posted them on its Web site, claymath.org. Meanwhile, Huai-Dong Cao of Lehigh University and Xi-Ping Zhu of Zhongshan University in Guangzhou, China, have published their own 318-page proof of both conjectures in The Asian Journal of Mathematics (www.ims.cuhk.edu.hk/).

Although these works were all hammered out in the midst of discussion and argument by experts, in workshops and lectures, they are about to receive even stricter scrutiny and perhaps crossfire. “Caution is appropriate,” said Dr. Kleiner, because the Poincaré conjecture is not just famous, but important.

James Carlson, president of the Clay Institute, said the appearance of these papers had started the clock ticking on a two-year waiting period mandated by the rules of the Clay Millennium Prize. After two years, he said, a committee will be appointed to recommend a winner or winners if it decides the proof has stood the test of time.

“There is nothing in the rules to prevent Perelman from receiving all or part of the prize,” Dr. Carlson said, saying that Dr. Perelman and Dr. Hamilton had obviously made the main contributions to the proof.

In a lecture at M.I.T. in 2003, Dr. Perelman described himself “in a way” as Dr. Hamilton’s disciple, although they had never worked together. Dr. Hamilton, who got his Ph.D. from Princeton in 1966, is too old to win the Fields medal, which is given only up to the age of 40, but he is slated to give the major address about the Poincaré conjecture in Madrid next week. He did not respond to requests for an interview.

Allowing that Dr. Perelman, should he win the Clay Prize, might refuse the honor, Dr. Carlson said the institute could decide instead to use award money to support Russian mathematicians, the Steklov Institute or even the Math Olympiad.

Dr. Anderson said that to some extent the new round of papers already represented a kind of peer review of Dr. Perelman’s work. “All these together make the case pretty clear,” he said. “The community accepts the validity of his work. It’s commendable that the community has gotten together.”

Correction: Aug. 21, 2006

An article in Science Times on Tuesday about a mathematical problem called the Poincaré conjecture misstated the year William Thurston, the author of a deeper conjecture that includes Poincaré’s, was awarded a Fields Medal. It was 1982, not 1986.

SEEDMAGAZINE.COM
APRIL 12, 2011

What is The Poincare Conjecture?

BY **STEPHEN ORNES** / AUGUST 25, 2006

Credit: Andrey Krasnov

In a 1904 paper, the French mathematician Jules Henri Poincaré stated that a sphere is a sphere is a sphere. You can punch, kick and throw it; you can inflate or deflate it; you can mold the sphere into another shape. But in the world of topology, no matter what you do to it, the resulting deformed, twisted and complicated form is still a sphere.

Also, you cannot poke a hole in it. You cannot, for example, turn your sphere into a donut. You cannot turn it into a coffee cup with a handle, frames for your eyeglasses or a key ring. You can stick your finger into it, but you can't actually puncture the surface or reach inside. If you break the surface in any way you've ventured into a different genus of topological objects.

Say you're walking down a street, and you encounter a strange and complicated shape whose surface sports peaks and valleys, mountains and molehills, but no holes. If you were a mathematician, you may want to study the way that functions behave on it. Poincaré's conjecture says that no matter what it looks like, it's a sphere. The conjecture gives mathematicians a short and easy way to identify a deformed blob as a sphere in disguise.

There is one more complication: When most people think of a sphere, they generally consider the space that a sphere occupies—a ping-pong ball, for example. When topologists talk about a sphere, they are talking exclusively about its surface.

A 1-sphere, for example, is the outside of a circle. A 2-sphere is the curved surface of what we consider a sphere. It is two-dimensional because, if you stand on the surface and look around, it appears that you are in a two-dimensional space. The surface of the Earth serves as a rough analogy—the world essentially appears flat when we stand on the ground and survey the horizon.

Mathematician solves Poincaré, rejects prize

Grigory Perelman solved the 100-year-old Poincaré conjecture, which won him the right to be the first mathematician ever to turn down the Field's Medal. Click here to read his story.

In its original form, however, the Poincaré conjecture concerned three-dimensional spheres (i.e., 3-spheres). These shapes are difficult, perhaps impossible, to visualize—the universe, for instance, is thought of as a 3-sphere. Even without being able to picture it, draw it or know that it exists, we can do math on a 3-sphere. We can calculate distances between points. Any system that can be characterized by three numbers automatically determines a three-dimensional shape. In baseball, for example, if you tally the numbers of runs, pitches and fouls for each inning of a game that doesn't go into extra innings, you have established nine data points in a three-dimensional space. With those nine points, you can make statements about the "shape" you have created.

Now imagine that this 3-sphere universe is distorted, wadded, dilated and deformed (but not punctured). If we lived on this deformed 3-sphere, you could feasibly walk across the Golden Gate Bridge and end up to Mars.

Or, you could think about it this way: You tie a lasso around your blob and tighten it until the string lies on the surface. If, for *every* different way you can tie the lasso, you can slip it off, then the blob is a sphere. The term for this is "simply connected." If it is possible to tie the lasso in such a way that it proves impossible to remove the lasso without breaking either the rope or the blob, it is not a sphere.

A donut, for example, is not "simply connected." If your lasso passes through the center of a donut, you cannot remove it without either altering the shape of the donut or cutting the rope. Though this is not its mathematically precise term, "breaking the donut" is absolutely, positively not allowed in the world of topology.

To illustrate Perelman's work, Jim Carlson, president of the Clay Mathematical Institute, draws a very complicated, closed squiggle. "This is really a circle, but it's a very wild circle," he says. (If his squiggle were an island, it would be an island ringed with fjords.)

"The idea is, in some sense, to apply heat to the shape and to allow the heat to simplify it. Take this very complicated wild circle, and imagine putting a little air hose in here and inflating it," he said, drawing a little box next to the squiggle. "It will dilate, and eventually it will achieve a round shape. Imagine a crinkled up balloon - you want to know what its real shape is, well blow it up with air, and then look at it. It achieves the simplest possible shape after you blow it up enough."

This notion of adding air, or heat, to a complicated shape was first developed by Richard Hamilton in the 1980s and is called Ricci Flow. Hamilton came close to solving the Poincaré conjecture, but he failed to successfully account for all singularities that may arise on the object. Singularities might be thought of as places where the fabric of the object is 'pinched.' Imagine that the balloon, for example, turned out to be shaped like a barbell. In that case, the two sides of the barbell would continue to inflate, while the connecting rod became thinner and thinner. As time goes on, it will not resemble a sphere. This connecting rod would be considered a singularity.

Perelman’s insight was to essentially ‘snip’ the rod. By utilizing abstract scissors, Perelman’s method allowed each side of the barbell to become its own sphere. The two resulting spheres would be topologically indistinguishable. Mathematicians refer to this process as ‘surgery’ on a 3-dimensional object.

Perelman’s use of surgery on these complicated surfaces was unprecedented and unexpected.

Slate

Who Cares About Poincaré?

Million-dollar math problem solved. So what?

By Jordan Ellenberg

Posted Friday, Aug. 18, 2006, at 11:59 AM ET

The *New York Times* recently reported that reclusive Russian geometer Grigory Perelman has apparently proved the century-old Poincaré conjecture. The *Times* calls Poincaré "a landmark not just of mathematics, but of human thought." But just *why* it's so significant is left a bit hazy. Big mathematical advances often generate the same kind of lofty but c ontent-free rhetoric found in political speeches about "the family." Like the family, math is a subject everyone agrees is very important without being able to specify exactly why.

I'm here to help. (With the Poincaré conjecture. As for the family, you're on your own.) Poincaré conjectured that three-dimensional shapes that share certain easy-to-check properties with spheres actually *are* spheres. What are these properties? My fellow geometer Christina Sormani describes the setup as follows:

That's zingier than anything the *Times* will run, but may still leave you without a clear picture of Perelman's theorem. Indeed, it's pretty hard to give an elementary account of the statement that Poincaré conjectured and that Perelman seems to have confirmed. (If that's what you're after, Sormani's home page links to a variety of expositions, including one in the form of a short story.) Instead, I'll try to explain why Perelman's theorem matters *without* explaining what it is.

The entities we study in science fall into two categories: those which can be

Slate

Who Cares About Poincaré?

classified in a way a human can understand, and those which are unclassifiably wild. Numbers are in the first class—you would agree that although you cannot *list* all the whole numbers, you have a good sense of what numbers are out there. Platonic solids are another good example. There are just five: the tetrahedron, the cube, the octahedron, the dodecahedron, and the icosahedron. End of story—you know them all. These mathematical objects behave something like the chemical elements, which are neatly classified by Mendeleev's periodic table. Many properties of an element are determined by its place in the table. For instance, we knew a lot about how metals like germanium and gallium would behave before they were actually discovered in nature!

In the second class are things like networks (in mathematical lingo, *graphs*) and beetles. There doesn't appear to be any nice, orderly structure on the set of all beetles, and we've got no way to predict what kinds of novel species will turn up. All we can do is observe some features that most beetles seem to share, most of the time. But there's no periodic table of beetles, and there probably couldn't be.

Mathematicians are much happier when a mathematical subject turns out to be of the first, more structured, type. We are much sadder when a subject turns out to be a variegated mass of beetles. (But have a look at Fields Medalist Timothy Gowers' beautiful essay "The Two Cultures of Mathematics" for a spirited defense of mathematical enterprises of the second sort.)

So, where do three-dimensional shapes, the subject of the Poincaré conjecture, fit in? To simplify, let's think about two-dimensional shapes first. These fall firmly in the "periodic table" category. The only such shapes are the surfaces of "doughnuts" with multiple holes. The number of holes is called the *genus* of the surface and plays the role that the atomic number does for chemical elements. (Here is a picture of the surfaces of genus 0, 1, 2, and 3.)

Fast, convenient, AND affordable
DELL
Save on select Wireless* All-in-Ones, Photo Printers and Lasers
Save Today
Limited time offers
Shop Now

Slate

Who Cares About Poincaré?

Geometer William Thurston (another Fields winner) made the daring conjecture that three-dimensional shapes, too, can be classified in a more complicated but equally structured way. Perelman has proved this conjecture, which has Poincaré as a straightforward corollary. That means, in turn, that we can think about proving general statements about three-dimensional geometry in a way that we can't hope to about beetles or graphs.

Perelman's work isn't important because of its applications. It won't help anyone build a bridge, aim a rocket, crack a code, or privatize Social Security. Mathematicians, no dummies, like to point out that, in some unspecified future, Perelman's theorem *might* pitch in to help with these problems in ways that aren't obvious now. But its real significance is like that of the fact that a times b is equal to b times a; it's a basic structural statement about how the world is organized. If you prefer order to chaos, that's something worth caring about.

HISTORY AND BACKGROUND

In the latter part of the nineteenth century, the French mathematician Henri Poincaré was studying the problem of whether the solar system is stable. Do the planets and asteroids in the solar system continue in regular orbits for all time, or will some of them be ejected into the far reaches of the galaxy or, alternatively, crash into the sun? In this work he was led to topology, a still new kind of mathematics related to geometry, and to the study of shapes (compact manifolds) of all dimensions.

The simplest such shape was the circle, or distorted versions of it such as the ellipse or something much wilder: lay a piece of string on the table, tie one end to the other to make a loop, and then move it around at random, making sure that the string does not touch itself. The next simplest shape is the two-sphere, which we find in nature as the idealized skin of an orange, the surface of a baseball, or the surface of the earth, and which we find in Greek geometry and philosophy as the "perfect shape." Again, there are distorted versions of the shape, such as the surface of an egg, as well as still wilder objects. Both the circle and the two-sphere can be described in words or in equations as the set of points at a fixed distance from a given point (the center). Thus it makes sense to talk about the three-sphere, the four-sphere, etc. These shapes are hard to visualize, since they naturally are contained in four-dimensional space, five-dimensional space, and so on, whereas we live in three-dimensional space. Nonetheless, with mathematical training, shapes in higher-dimensional spaces can be studied just as well as shapes in dimensions two and three.

In topology, two shapes are considered the same if the points of one correspond to the points of another in a continuous way. Thus the circle, the ellipse, and the wild piece of string are considered the same. This is much like what happens in the geometry of Euclid. Suppose that one shape can be moved, without changing lengths or angles, onto another shape. Then the two shapes are considered the same (think of congruent triangles). A round, perfect two-sphere, like the surface of a ping-pong ball, is topologically the same as the surface of an egg.

In 1904 Poincaré asked whether a three-dimensional shape that satisfies the "simple connectivity test" is the same, topologically, as the ordinary round three-sphere. The round three-sphere is the set of points equidistant from a given point in four-dimensional space. His test is something that can be performed by an imaginary being who lives inside the three-dimensional shape and cannot see it from "outside." The test is that every loop in the shape can be drawn back to the point of departure without leaving the shape. This can be done for the two-sphere and the three-sphere. But it cannot be done for the surface of a doughnut, where a loop may get stuck around the hole in the doughnut.

The question raised became known as the Poincaré conjecture. Over the years, many outstanding mathematicians tried to solve it—Poincaré himself, Whitehead, Bing, Papakirioukopolos, Stallings, and others. While their efforts frequently led to the creation of significant new mathematics, each time a flaw was found in the proof. In 1961 came astonishing news. Stephen Smale, then of the University of California at Berkeley (now at the City University of Hong Kong) proved that the analogue of the Poincaré conjecture was true for spheres of five or more dimensions. The higher-dimensional version of the conjecture required a more stringent version of Poincaré's test; it asks whether a so-called homotopy sphere is a true sphere. Smale's theorem was an achievement of extraordinary proportions. It did not, however, answer Poincaré's original question. The search for an answer became all the more alluring.

Smale's theorem suggested that the theory of spheres of dimensions three and four was unlike the theory of spheres in higher dimension. This notion was confirmed a decade later, when Michael Freedman, then at the University of California, San Diego, now of Microsoft Research Station Q, announced a proof of the Poincaré conjecture in dimension four. His work used techniques quite different from those of Smale. Freedman also gave a classification, or kind of species list, of all simply connected four-dimensional manifolds.

Both Smale (in 1966) and Freedman (in 1986) received Fields medals for their work.

There remained the original conjecture of Poincaré in dimension three. It seemed to be the most difficult of all, as the continuing series of failed efforts, both to prove and to disprove it, showed. In the meantime, however, there came three developments that would play crucial roles in Perelman's solution of the conjecture.

Geometrization

The first of these developments was William Thurston's geometrization conjecture. It laid out a program for understanding all three-dimensional shapes in a coherent way, much as had been done for two-dimensional shapes in the latter half of the nineteenth century. According to Thurston, three-dimensional shapes could be broken down into pieces governed by one of eight geometries, somewhat as a molecule can be broken into its constituent, much simpler atoms. This is the origin of the name, "geometrization conjecture."

A remarkable feature of the geometrization conjecture was that it implied the Poincaré conjecture as a special case. Such a bold assertion was accordingly thought to be far, far out of reach—perhaps a subject of research for the twenty-second century. Nonetheless, in an imaginative *tour de force* that drew on many fields of mathematics, Thurston was able to prove the geometrization conjecture for a wide class of shapes (Haken manifolds) that have a sufficient degree of complexity. While these methods did not apply to the three-sphere, Thurston's work shed new light on the central role of Poincaré's conjecture and placed it in a far broader mathematical context.

Limits of spaces

The second current of ideas did not appear to have a connection with the Poincaré conjecture until much later. While technical in nature, the work, in which the names of Cheeger and Perelman figure prominently, has to do with how one can take limits of geometric shapes, just as we learned to take limits in beginning calculus class. Think of Zeno and his paradox: you walk half the distance from where you are standing to the wall of your living room. Then you walk half the remaining distance. And so on. With each step you get closer to the wall. The wall is your "limiting position," but you never reach it in a finite number of steps. Now imagine a shape changing with time. With each "step" it changes shape, but can nonetheless be a "nice" shape at each step— smooth, as the mathematicians say. For the limiting shape the situation is different. It may be nice and smooth, or it may have special points that are different from all the others, that is, singular points, or “singularities.” Imagine a Y-shaped piece of tubing that is collapsing: as time increases, the diameter of the tube gets smaller and smaller. Imagine further that one second after the tube begins its collapse, the diameter has gone to zero. Now the shape is different: it is a Y shape of infinitely thin wire. The point where the arms of the Y meet is different from all the others. It is the singular point of this shape. The kinds of shapes that can occur as limits are called Aleksandrov spaces, named after the Russian mathematician A. D. Aleksandrov who initiated and developed their theory.

Differential equations

The third development concerns differential equations. These equations involve rates of change in the unknown quantities of the equation, e.g., the rate of change of the position of an apple as it falls from a tree towards the earth's center. Differential equations are expressed in the language of calculus, which Isaac Newton invented in the 1680s in order to explain how material bodies (apples, the moon, and so on) move under the influence of an external force. Nowadays physicists use differential equations to study a great range of phenomena: the motion of galaxies and the stars within them, the flow of air and water, the propagation of sound and light, the conduction of heat, and even the creation, interaction, and annihilation of elementary particles such as electrons, protons, and quarks.

In our story, conduction of heat and change of temperature play a special role. This kind of physics was first treated mathematically by Joseph Fourier in his 1822 book, *Théorie Analytique de la Chaleur.* The differential equation that governs change of temperature is called the heat equation. It has the remarkable property that as time increases, irregularities in the distribution of temperature decrease.

Differential equations apply to geometric and topological problems as well as to physical ones. But one studies not the rate at which temperature changes, but rather the rate of change in some geometric quantity as it relates to other quantities such as curvature. A piece of paper lying on the table has curvature zero. A sphere has positive curvature. The curvature is a large number for a small sphere, but is a small number for a large sphere such as the surface of the earth. Indeed, the curvature of the earth is so small that its surface has sometimes mistakenly been thought to be flat. For an example of negative curvature, think of a point on the bell of a trumpet. In some directions the metal bends away from your eye; in others it bends towards it.

An early landmark in the application of differential equations to geometric problems was the 1963 paper of J. Eells and J. Sampson. The authors introduced the "harmonic map equation," a kind of nonlinear version of Fourier's heat equation. It proved to be a powerful tool for the solution of geometric and topological problems. There are now many important nonlinear heat equations—the equations for mean curvature flow, scalar curvature flow, and Ricci flow.

Also notable is the Yang-Mills equation, which came into mathematics from the physics of quantum fields. In 1983 this equation was used to establish very strong restrictions on the topology of four-dimensional shapes on which it was possible to do calculus [D]. These results helped renew hopes of obtaining other strong geometric results from analytic arguments—that is, from calculus and differential equations. Optimism for such applications had been tempered to some extent by the examples of René Thom (on cycles not representable by smooth submanifolds) and Milnor (on diffeomorphisms of the six-sphere).

Ricci flow

The differential equation that was to play a key role in solving the Poincaré conjecture is the Ricci flow equation. It was discovered two times, independently. In physics, the equation originated with the thesis of Friedan [F, 1985], although it was perhaps implicit in the work of Honerkamp [Ho, 1972]. In mathematics it originated with the 1982 paper of Richard Hamilton [Ha1]. The physicists were working on the renormalization group of quantum field theory, while Hamilton was interested in geometric applications of the Ricci flow equation itself. Hamilton, now at Columbia University, was then at Cornell University.

On the left-hand side of the Ricci flow equation is a quantity that expresses how the geometry changes with time—the derivative of the metric tensor, as the mathematicians like to say. On the right-hand side is the Ricci tensor, a measure of the extent to which the shape is curved. The Ricci tensor, based on Riemann's theory of geometry (1854), also appears in Einstein's equations for general relativity (1915). Those equations govern the interaction of matter, energy, curvature of space, and the motion of material bodies.

The Ricci flow equation is the analogue, in the geometric context, of Fourier's heat equation. The idea, *grosso modo*, for its application to geometry is that, just as Fourier's heat equation disperses temperature, the Ricci flow equation disperses curvature. Thus, even if a shape was irregular and distorted, Ricci flow would gradually remove these anomalies, resulting in a very regular shape whose topological nature was evident. Indeed, in 1982 Hamilton showed that for positively curved, simply connected shapes of dimension three (compact three-manifolds) the Ricci flow transforms the shape into one that is ever more like the round three-sphere. In the long run, it becomes almost indistinguishable from this perfect, ideal shape. When the curvature is not strictly positive, however, solutions of the Ricci flow equation behave in a much more complicated way. This is because the equation is nonlinear. While parts of the shape may evolve towards a smoother, more regular state, other parts might develop singularities. This richer behavior posed serious difficulties. But it also held promise: it was conceivable that the formation of singularities could reveal Thurston's decomposition of a shape into its constituent geometric atoms.

Richard Hamilton

Hamilton was the driving force in developing the theory of Ricci flow in mathematics, both conceptually and technically. Among his many notable results is his 1999 paper [Ha2], which showed that in a Ricci flow, the curvature is pushed towards the positive near a singularity. In that paper Hamilton also made use of the collapsing theory [C-G] mentioned earlier. Another result [Ha3], which played a crucial role in Perelman's proof, was the Hamilton Harnack inequality, which generalized to positive Ricci flows a result of Peter Li and Shing-Tung Yau for positive solutions of Fourier's heat equation.

Hamilton had established the Ricci flow equation as a tool with the potential to resolve both conjectures as well as other geometric problems. Nevertheless, serious obstacles barred the way to a proof of the Poincaré conjecture. Notable among these obstacles was lack of an adequate understanding of the formation of singularities in Ricci flow, akin to the formation of black holes in the evolution of the cosmos. Indeed, it was not at all clear how or if formation of singularities could be understood. Despite the new front opened by Hamilton, and despite continued work by others using traditional topological tools for either a proof or a disproof, progress on the conjectures came to a standstill.

Such was the state of affairs in 2000, when John Milnor wrote an article describing the Poincaré conjecture and the many attempts to solve it. At that writing, it was not clear whether the conjecture was true or false, and it was not clear which method might decide the issue. Analytic methods (differential equations) were mentioned in a later version (2004). See [M1] and [M2].

Perelman announces a solution of the Poincaré conjecture

It was thus a huge surprise when Grigoriy Perelman announced, in a series of preprints posted on ArXiv.org in 2002 and 2003, a solution not only of the Poincaré conjecture, but also of Thurston's geometrization conjecture [P1, P2, P3].

The core of Perelman's method of proof is the theory of Ricci flow. To its applications in topology he brought not only great technical virtuosity, but also new ideas. One was to combine collapsing theory in Riemannian geometry with Ricci flow to give an understanding of the parts of the shape that were collapsing onto a lower-dimensional space. Another was the introduction of a new quantity, the entropy, which instead of measuring disorder at the atomic level, as in the classical theory of heat exchange, measures disorder in the global geometry of the space. Perelman's entropy, like the thermodynamic entropy, is increasing in time: there is no turning back. Using his entropy function and a related local version (the L-length functional), Perelman was able to understand the nature of the singularities that formed under Ricci flow. There were just a few kinds, and one could write down simple models of their formation. This was a breakthrough of first importance.

Once the simple models of singularities were understood, it was clear how to cut out the parts of the shape near them as to continue the Ricci flow past the times at which they would otherwise form. With these results in hand, Perelman showed that the formation times of the singularities could not run into Zeno's wall: imagine a singularity that occurs after one second, then after half a second more, then after a quarter of a second more, and so on. If this were to occur, the "wall," which one would reach two seconds after departure, would correspond to a time at which the mathematics of Ricci flow would cease to hold. The proof would be unattainable. But with this new mathematics in hand, attainable it was.

The posting of Perelman's preprints and his subsequent talks at MIT, SUNY–Stony Brook, Princeton, and the University of Pennsylvania set off a worldwide effort to understand and verify his groundbreaking work. In the US, Bruce Kleiner and John Lott wrote a set of detailed notes on Perelman's work. These were posted online as the verification effort proceeded. A final version was posted to ArXiv.org in May 2006, and the refereed article appeared in *Geometry and Topology* in 2008. This was the first time that work on a problem of such importance was facilitated via a public website. John Morgan and Gang Tian wrote a book-long exposition of Perelman's proof, posted on ArXiv.org in July of 2006, and published by the American Mathematical Society in CMI's monograph series (August 2007). These expositions, those by other teams, and, importantly, the multi-year scrutiny of the mathematical community, provided the needed verification. Perelman had solved the Poincaré conjecture. After a century's wait, it was settled!

Among other articles that appeared following Perelman's work is a paper in the *Asian Journal of Mathematics*, posted on ArXiv.org in June of 2006 by the American-Chinese team, Huai-Dong Cao (Lehigh University) and Xi-Ping Zhu (Zhongshan University). Another is a paper by the European group of Bessières, Besson, Boileau, Maillot, and Porti, posted on ArXiv.org in June of 2007. It was accepted for publication by *Inventiones Mathematicae* in October of 2009. It gives an alternative approach to the last step in Perelman's proof of the geometrization conjecture.

Perelman's proof of the Poincaré and geometrization conjectures is a major mathematical advance. His ideas and methods have already found new applications in analysis and geometry; surely the future will bring many more.

— JC, March 18, 2010 (corrections, 3/19/2010)

References

[C-G] J. Cheeger and M. Gromov, Collapsing Riemannian manifolds while keeping their curvature bounded. I and II, J. Differential Geom. Volume 23, Number 3 (1986); Volume 32, Number 1 (1990), 269-298.

[D] S.K. Donaldson. An application of gauge theory to four-dimensional topology. J. Differential Geom., 18, (1983), 279–315.

[F] D. Friedan, Nonlinear Models in 2 + epsilon Dimensions, Annals of Physics 163, 318-419 (1985)

[Ha1] R. Hamilton, Three-manifolds with positive Ricci curvature, Journal of Differential Geometry, vol. 17:255-306 (1982)

[Ha2] R. Hamilton, Non-singular solutions of the Ricci flow on three-manifolds, Comm. Anal. Geom. 7(4): 695-729 (1999)

[Ha3] R. Hamilton, The Harnack estimate for Ricci flow, Journal of Differential Geometry, vol. 37:225-243 (1993)

[Ho] J. Honerkamp, (CERN), Chiral multiloops, Nucl. Phys. B36:130-140 (1972)

[M1] J. Milnor, The Poincaré Conjecture (2000) www.claymath.org/millennium/Poincare_Conjecture/poincare_2000.pdf

[M2] J. Milnor, The Poincaré Conjecture, in *The Millennium Prize Problems,* J. Carlson, A. Jaffe, A. Wiles, eds, AMS (2004) www.claymath.org/millennium/Poincare_Conjecture/poincare.pdf

[P1] G. Perelman, The entropy formula for the Ricci flow and its geometric applications, arXiv.org, November 11, 2002

[P2] G. Perelman, Ricci flow with surgery on three-manifolds, arXiv.org, March 10, 2003

[P3] G. Perelman, Finite extinction time for the solutions to the Ricci flow on certain three-manifolds, arXiv.org, July 17, 2003

###

The Vector Potential
Aharonov-Bohm Effect
Gauge Transformations

Feynman’s Paradox

A paradox is a situation which gives one answer when analyzed one way, and a different answer when analyzed another way, so that we are left in somewhat of a quandary as to actually what would happen. Of course, in physics there are never any real paradoxes because there is one correct answer; at least we believe that nature will act in only one way (and that is the *right way*, naturally). So a paradox in physics is only a confusion in our understanding.

act in only one way (and that is the *right way*, naturally). So in physics a paradox is only a confusion in our own understanding. Here is our paradox.

Imagine that we construct a device like that shown in Fig. 17–5. There is a thin, circular plastic disc supported on a concentric shaft with excellent bearings, so that it is quite free to rotate. On the disc is a coil of wire in the form of a short solenoid concentric with the axis of rotation. This solenoid carries a steady current I provided by a small battery, also mounted on the disc. Near the edge of the disc and spaced uniformly around its circumference are a number of small metal spheres insulated from each other and from the solenoid by the plastic material of the disc. Each of these small conducting spheres is charged with the same electrostatic charge Q. Everything is quite stationary, and the disc is at rest. Suppose now that by some accident—or by prearrangement—the current in the solenoid is interrupted, without, however, any intervention from the outside. So long as the current continued, there was a magnetic flux through the solenoid more or less parallel to the axis of the disc. When the current is interrupted, this flux must go to zero. There will, therefore, be an electric field induced which will circulate around in circles centered at the axis. The charged spheres on the perimeter of the disc will all experience an electric field tangential to the perimeter of the disc. This electric force is in the same sense for all the charges and so will result in a net torque on the disc. From these arguments we would expect that as the current in the solenoid disappears, the disc would begin to rotate. If we knew the moment of inertia of the disc, the current in the solenoid, and the charges on the small spheres, we could compute the resulting angular velocity.

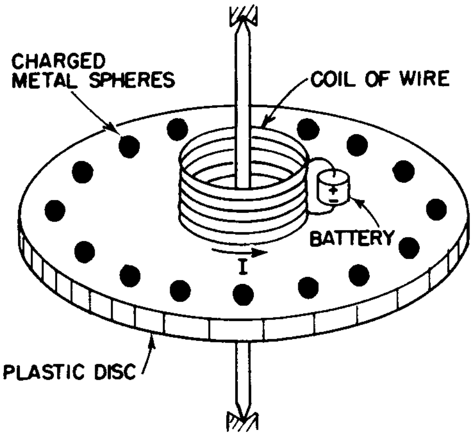


Fig. 17–5. Will the disc rotate if the current I is stopped?

But we could also make a different argument. Using the principle of the conservation of angular momentum, we could say that the angular momentum of the disc with all its equipment is initially zero, and so the angular momentum of the assembly should remain zero. There should be no rotation when the current is stopped. Which argument is correct? Will the disc rotate or will it not? We will leave this question for you to think about.

We should warn you that the correct answer does not depend on any nonessential feature, such as the asymmetric position of a battery, for example. In fact, you can imagine an ideal situation such as the following: The solenoid is made of superconducting wire through which there is a current. After the disc has been carefully placed at rest, the temperature of the solenoid is allowed to rise slowly. When the temperature of the wire reaches the transition temperature between superconductivity and normal conductivity, the current in the solenoid will be brought to zero by the resistance of the wire. The flux will, as before, fall to zero, and there will be an electric field around the axis. We should also warn you that the solution is not easy, nor is it a trick. When you figure it out, you will have discovered an important principle of electromagnetism.

17–5 Alternating-current generator

In the remainder of this chapter we apply the principles of Section 17–1 to analyze a number of the phenomena discussed in Chapter 16. We first look in more detail at the alternating-current generator. Such a generator consists basically of a coil of wire rotating in a uniform magnetic field. The same result can also be achieved by a fixed coil in a magnetic field whose direction rotates in the manner described in the last chapter. We will consider only the former case. Suppose we have a circular coil of wire which can be turned on an axis along one of its diameters. Let this coil be located in a uniform magnetic field perpendicular to the axis of rotation, as in Fig. 17–6. We also imagine that the two ends of the coil are brought to external connections through some kind of sliding contacts.

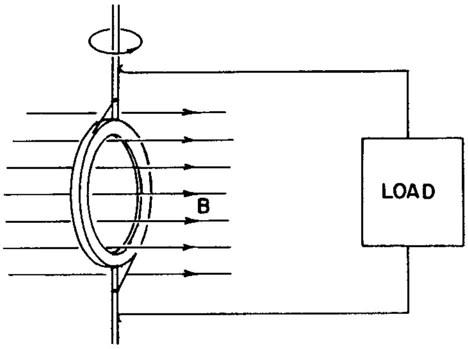


Fig. 17–6. A coil of wire rotating in a uniform magnetic field—the basic idea of the ac generator.

Due to the rotation of the coil, the magnetic flux through it will be changing. The circuit of the coil will therefore have an emf in it. Let S be the area of the coil and θ the angle between the magnetic field and the normal to the plane of the coil.*

* Now that we are using the letter A for the vector potential, we prefer to let S stand for a Surface area.

The Vector Potential

EM Field Carries Momentum

A Turns into Photons

Generalized Momentum π

Particle Momentum p

Field Momentum $(e/c)A$

$$\pi = p - (e/c)\, A$$

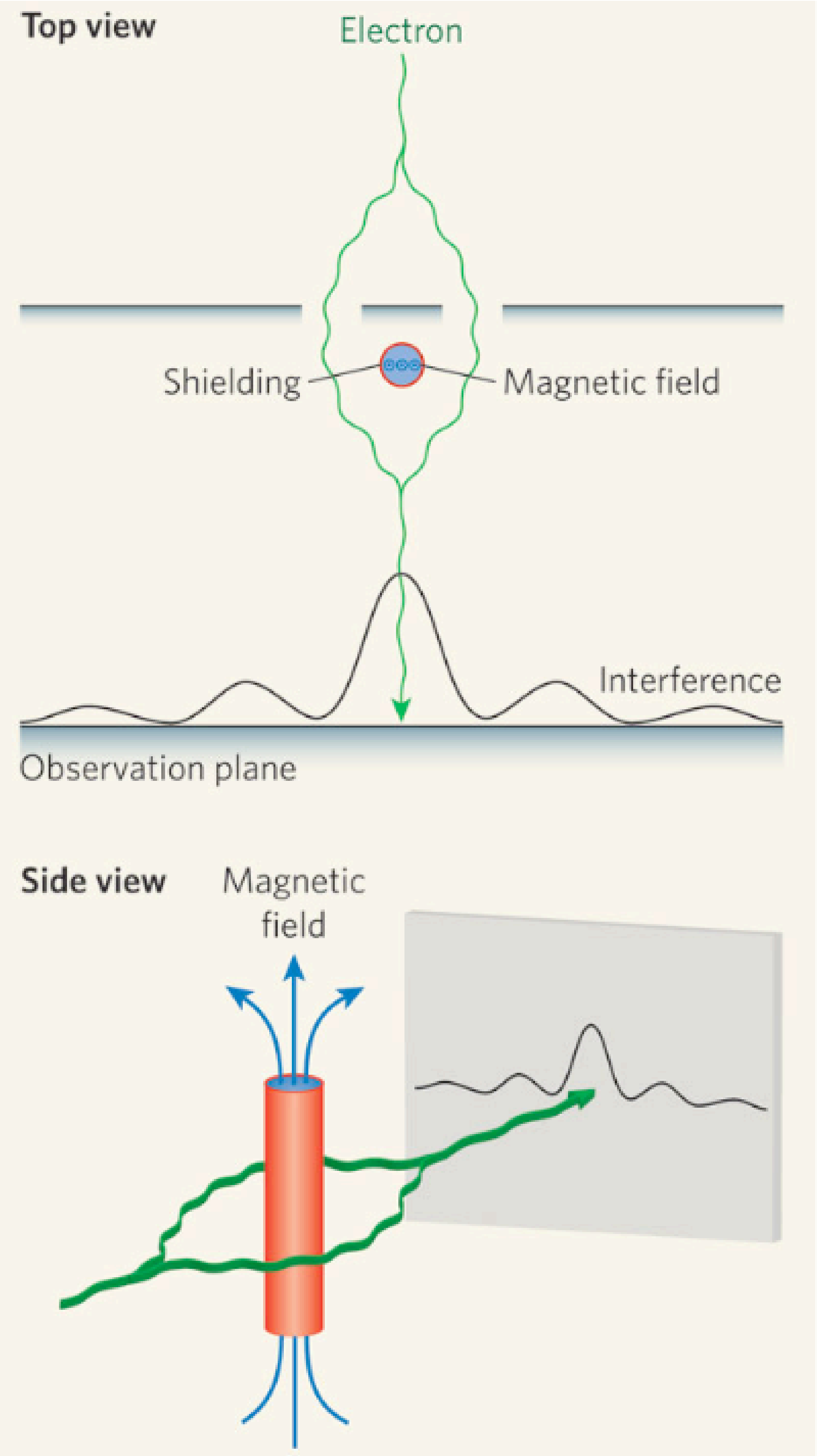
$$E = \pi^2 / 2m$$

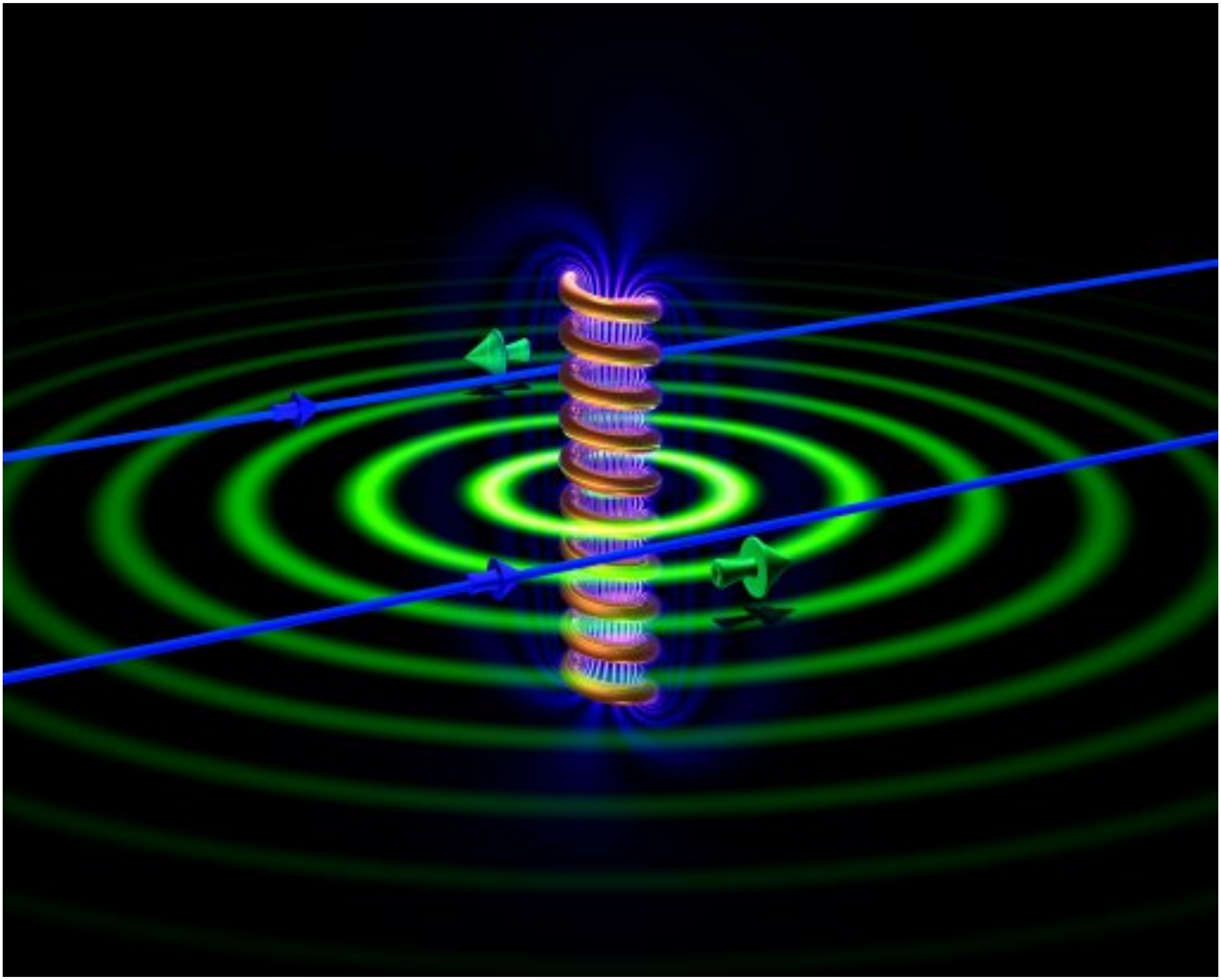
$$2mE = p^2 - 2\,(e/c)\, p \cdot A + (e/c)^2\, A^2$$

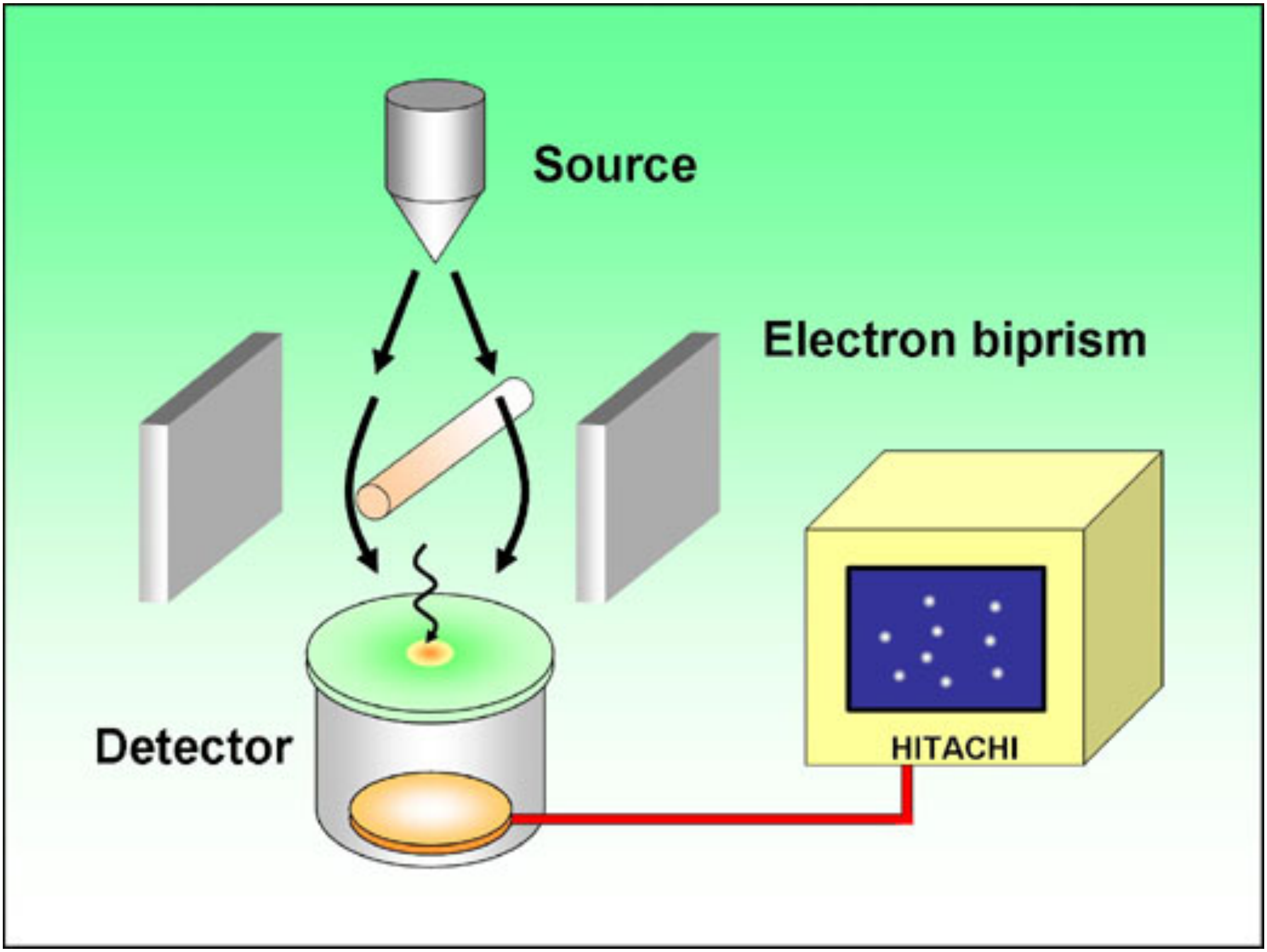
The Aharonov-Bohm Effect

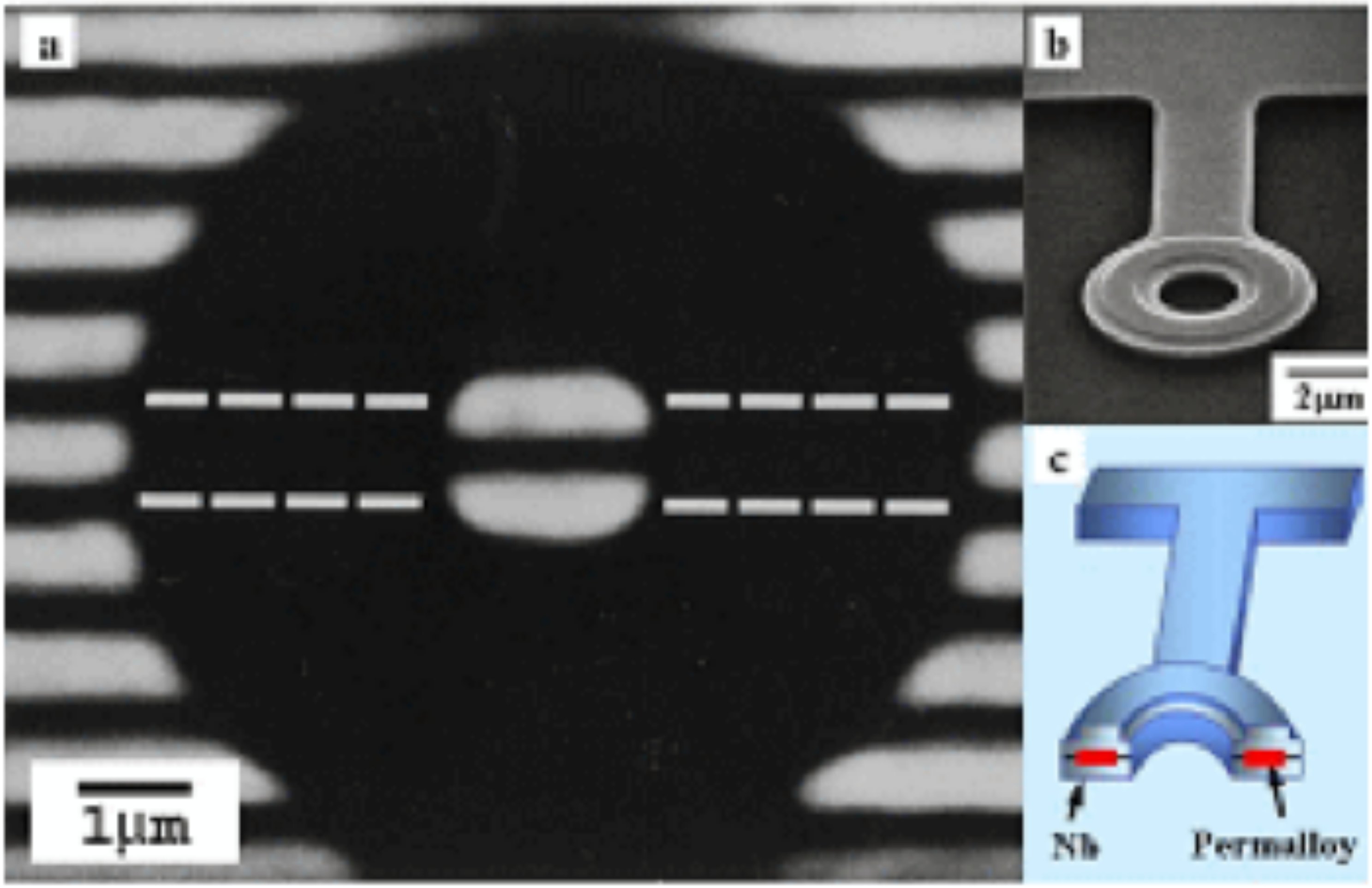
An electron moving in a region where E and B are zero, but A is not exhibits physical effects.

Therefore A is real whereas E and B are not.


Top view
Electron
Shielding
Magnetic field
Interference
Observation plane
Side view
Magnetic
field




Source
Electron biprism
Detector
HITACHI


a
1μm
b
2μm
c
Nb
Permalloy

The advantages are much less clear for magnetostatics. The integral for $\boldsymbol{A}$ is already a vector integral:

$$\boldsymbol{A}(1) = \frac{1}{4\pi\epsilon_0 c^2}\int \frac{\boldsymbol{j}(2)\,dV_2}{r_{12}}, \tag{15.24}$$

which is, of course, three integrals. Also, when we take the curl of $\boldsymbol{A}$ to get $\boldsymbol{B}$, we have six derivatives to do and combine by pairs. It is not immediately obvious whether in most problems this procedure is really any easier than computing $\boldsymbol{B}$ directly from

$$\boldsymbol{B}(1) = \frac{1}{4\pi\epsilon_0 c^2}\int \frac{\boldsymbol{j}(2)\times \boldsymbol{e}_{12}}{r_{12}^2}\,dV_2. \tag{15.25}$$

Using the vector potential is often more difficult for simple problems for the following reason. Suppose we are interested only in the magnetic field $\boldsymbol{B}$ at one point, and that the problem has some nice symmetry—say we want the field at a point on the axis of a ring of current. Because of the symmetry, we can easily get $\boldsymbol{B}$ by doing the integral of Eq. (15.25). If, however, we were to find $\boldsymbol{A}$ first, we would have to compute $\boldsymbol{B}$ from *derivatives* of $\boldsymbol{A}$, so we must know what $\boldsymbol{A}$ is at all points in the *neighborhood* of the point of interest. And most of these points are off the axis of symmetry, so the integral for $\boldsymbol{A}$ gets complicated. In the ring problem, for example, we would need to use elliptic integrals. In such problems, $\boldsymbol{A}$ is clearly not very useful. It is true that in many complex problems it is easier to work with $\boldsymbol{A}$, but it would be hard to argue that this ease of technique would justify making you learn about one more vector field.

We have introduced $\boldsymbol{A}$ because it *does* have an important physical significance. Not only is it related to the energies of currents, as we saw in the last section, but it is also a "real" physical field in the sense that we described above. In classical mechanics it is clear that we can write the force on a particle as

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v}\times\boldsymbol{B}), \tag{15.26}$$

so that, given the forces, everything about the motion is determined. In any region where $\boldsymbol{B} = 0$ even if $\boldsymbol{A}$ is not zero, such as outside a solenoid, there is no discernible effect of $\boldsymbol{A}$. Therefore for a long time it was believed that $\boldsymbol{A}$ was not a "real" field. It turns out, however, that there are phenomena involving quantum mechanics which show that the field $\boldsymbol{A}$ is in fact a "real" field in the sense we have defined it. In the next section we will show you how that works.

15-5 The vector potential and quantum mechanics

There are many changes in what concepts are important when we go from classical to quantum mechanics. We have already discussed some of them in Vol. I. In particular, the force concept gradually fades away, while the concepts of energy and momentum become of paramount importance. You remember that instead of particle motions, one deals with probability amplitudes which vary in space and time. In these amplitudes there are wavelengths related to momenta, and frequencies related to energies. The momenta and energies, which determine the phases of wave functions, are therefore the important quantities in quantum mechanics. Instead of forces, we deal with the way interactions change the wavelength of the waves. The idea of a force becomes quite secondary—if it is there at all. When people talk about nuclear forces, for example, what they usually analyze and work with are the energies of interaction of two nucleons, and not the force between them. Nobody ever differentiates the energy to find out what the force looks like. In this section we want to describe how the vector and scalar potentials enter into quantum mechanics. It is, in fact, just because momentum and energy play a central role in quantum mechanics that $\boldsymbol{A}$ and ϕ provide the most direct way of introducing electromagnetic effects into quantum descriptions.

We must review a little how quantum mechanics works. We will consider again the imaginary experiment described in Chapter 37 of Vol. I, in which elec-

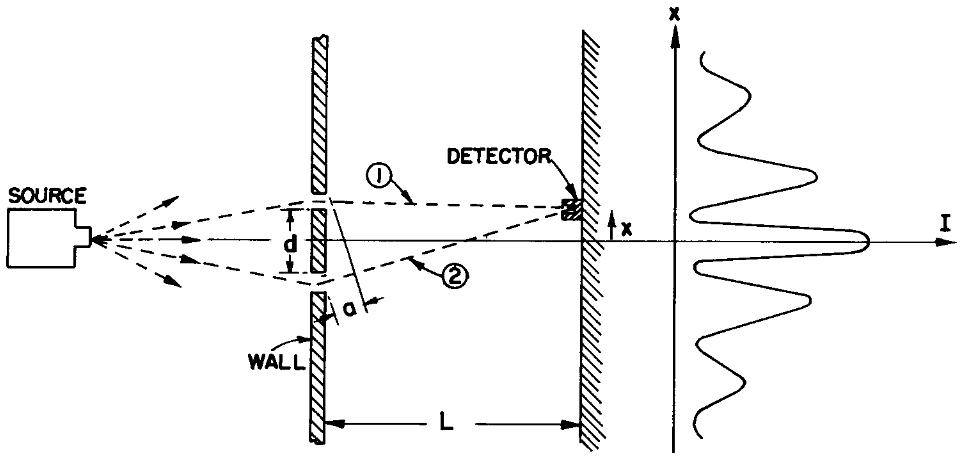


Fig. 15–5. An interference experiment with electrons (see also Chapter 37 of Vol. I).

trons are diffracted by two slits. The arrangement is shown again in Fig. 15–5. Electrons, all of nearly the same energy, leave the source and travel toward a wall with two narrow slits. Beyond the wall is a "backstop" with a movable detector. The detector measures the rate, which we call I, at which electrons arrive at a small region of the backstop at the distance x from the axis of symmetry. The rate is proportional to the probability that an individual electron that leaves the source will reach that region of the backstop. This probability has the complicated-looking distribution shown in the figure, which we understand as due to the interference of two amplitudes, one from each slit. The interference of the two amplitudes depends on their phase difference. That is, if the amplitudes are $C_1e^{i\Phi_1}$ and $C_2e^{i\Phi_2}$, the phase difference $\delta = \Phi_1 - \Phi_2$ determines their interference pattern [see Eq. (29.12) in Vol. I]. If the distance between the screen and the slits is L, and if the difference in the path lengths for electrons going through the two slits is a, as shown in the figure, then the phase difference of the two waves is given by

$$\delta = \frac{a}{\bar{\lambda}}. \tag{15.27}$$

As usual, we let $\bar{\lambda} = \lambda/2\pi$, where λ is the wavelength of the space variation of the probability amplitude. For simplicity, we will consider only values of x much less than L; then we can set

$$a = \frac{x}{L}d$$

and

$$\delta = \frac{x}{L}\frac{d}{\bar{\lambda}}. \tag{15.28}$$

When x is zero, δ is zero; the waves are in phase, and the probability has a maximum. When δ is π, the waves are out of phase, they interfere destructively, and the probability is a minimum. So we get the wavy function for the electron intensity.

Now we would like to state the law that for quantum mechanics replaces the force law $\boldsymbol{F} = q\boldsymbol{v} \times \boldsymbol{B}$. It will be the law that determines the behavior of quantum-mechanical particles in an electromagnetic field. Since what happens is determined by amplitudes, the law must tell us how the magnetic influences affect the amplitudes; we are no longer dealing with the acceleration of a particle. The law is the following: the phase of the amplitude to arrive via any trajectory is changed by the presence of a magnetic field by an amount equal to the integral of the vector potential along the whole trajectory times the charge of the particle over Planck's constant. That is,

$$\text{Magnetic change in phase} = \frac{q}{\hbar}\int_{\text{trajectory}} \boldsymbol{A}\cdot d\boldsymbol{s}. \tag{15.29}$$

If there were no magnetic field there would be a certain phase of arrival. If there is a magnetic field anywhere, the phase of the arriving wave is increased by the integral in Eq. (15.29).

Although we will not need to use it for our present discussion, we mention that the effect of an electrostatic field is to produce a phase change given by the *negative* of the *time* integral of the scalar potential ϕ:

$$\text{Electric change in phase} = -\frac{q}{\hbar}\int \phi\, dt.$$

These two expressions are correct not only for static fields, but together give the correct result for *any* electromagnetic field, static or dynamic. This is the law that replaces $\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})$. We want now, however, to consider only a static magnetic field.

Suppose that there is a magnetic field present in the two-slit experiment. We want to ask for the phase of arrival at the screen of the two waves whose paths pass through the two slits. Their interference determines where the maxima in the probability will be. We may call Φ_1 the phase of the wave along trajectory (1). If $\Phi_1(B = 0)$ is the phase without the magnetic field, then when the field is turned on the phase will be

$$\Phi_1 = \Phi_1(B = 0) + \frac{q}{\hbar}\int_{(1)} \boldsymbol{A} \cdot d\boldsymbol{s}. \tag{15.30}$$

Similarly, the phase for trajectory (2) is

$$\Phi_2 = \Phi_2(B = 0) + \frac{q}{\hbar}\int_{(2)} \boldsymbol{A} \cdot d\boldsymbol{s}. \tag{15.31}$$

The interference of the waves at the detector depends on the phase difference

$$\delta = \Phi_1(B = 0) - \Phi_2(B = 0) + \frac{q}{\hbar}\int_{(1)} \boldsymbol{A} \cdot d\boldsymbol{s} - \frac{q}{\hbar}\int_{(2)} \boldsymbol{A} \cdot d\boldsymbol{s}. \tag{15.32}$$

The no-field difference we will call $\delta(B = 0)$; it is just the phase difference we have calculated above in Eq. (15.28). Also, we notice that the two integrals can be written as *one* integral that goes forward along (1) and back along (2); we call this the closed path (1–2). So we have

$$\delta = \delta(B = 0) + \frac{q}{\hbar}\oint_{(1-2)} \boldsymbol{A} \cdot d\boldsymbol{s}. \tag{15.33}$$

This equation tells us how the electron motion is changed by the magnetic field; with it we can find the new positions of the intensity maxima and minima at the backstop.

Before we do that, however, we want to raise the following interesting and important point. You remember that the vector potential function has some arbitrariness. Two different vector potential functions $\boldsymbol{A}$ and $\boldsymbol{A}'$ whose difference is the gradient of some scalar function $\boldsymbol{\nabla}\psi$, both represent the same magnetic field, since the curl of a gradient is zero. They give, therefore, the same classical force $q\boldsymbol{v} \times \boldsymbol{B}$. If in quantum mechanics the effects depend on the vector potential, *which* of the many possible $\boldsymbol{A}$-functions is correct?

The answer is that the same arbitrariness in $\boldsymbol{A}$ continues to exist for quantum mechanics. If in Eq. (15.33) we change $\boldsymbol{A}$ to $\boldsymbol{A}' = \boldsymbol{A} + \boldsymbol{\nabla}\psi$, the integral on $\boldsymbol{A}$ becomes

$$\oint_{(1-2)} \boldsymbol{A}' \cdot d\boldsymbol{s} = \oint_{(1-2)} \boldsymbol{A} \cdot d\boldsymbol{s} + \oint_{(1-2)} \boldsymbol{\nabla}\psi \cdot d\boldsymbol{s}.$$

The integral of $\boldsymbol{\nabla}\psi$ is around the *closed* path (1–2), but the integral of the tangential component of a gradient on a closed path is always zero, by Stokes' theorem. Therefore both $\boldsymbol{A}$ and $\boldsymbol{A}'$ give the same phase differences and the same quantum-mechanical interference effects. In both classical and quantum theory it is only the curl of $\boldsymbol{A}$ that matters; any choice of the function of $\boldsymbol{A}$ which has the correct curl gives the correct physics.

The same conclusion is evident if we use the results of Section 14–1. There we found that the line integral of $\boldsymbol{A}$ around a closed path is the flux of $\boldsymbol{B}$ through the path, which here is the flux between paths (1) and (2). Equation (15.33) can, if we wish, be written as

$$\delta = \delta(B = 0) + \frac{q}{\hbar}\,[\text{flux of } \boldsymbol{B} \text{ between (1) and (2)}], \tag{15.34}$$

where by the flux of $\boldsymbol{B}$ we mean, as usual, the surface integral of the normal component of $\boldsymbol{B}$. The result depends only on $\boldsymbol{B}$, and therefore only on the curl of $\boldsymbol{A}$.

Now because we can write the result in terms of $\boldsymbol{B}$ as well as in terms of $\boldsymbol{A}$, you might be inclined to think that the $\boldsymbol{B}$ holds its own as a "real" field and that the $\boldsymbol{A}$ can still be thought of as an artificial construction. But the definition of "real" field that we originally proposed was based on the idea that a "real" field would not act on a particle from a distance. We can, however, give an example in which $\boldsymbol{B}$ is zero—or at least arbitrarily small—at any place where there is some chance to find the particles, so that it is not possible to think of it acting *directly* on them.

You remember that for a long solenoid carrying an electric current there is a $\boldsymbol{B}$-field inside but none outside, while there is lots of $\boldsymbol{A}$ circulating around outside, as shown in Fig. 15–6. If we arrange a situation in which electrons are to be found only *outside* of the solenoid—only where there is $\boldsymbol{A}$—there will still be an influence on the motion, according to Eq. (15.33). Classically, that is impossible. Classically, the force depends only on $\boldsymbol{B}$; in order to know that the solenoid is carrying current, the particle must go through it. But quantum-mechanically you can find out that there is a magnetic field inside the solenoid by going *around* it—without ever going close to it!

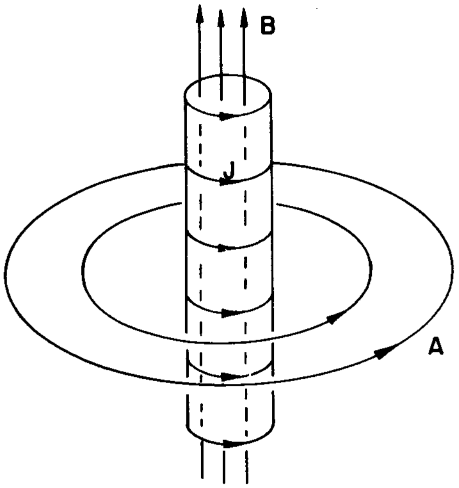


Fig. 15–6. The magnetic field and vector potential of a long solenoid.

Suppose that we put a very long solenoid of small diameter just behind the wall and between the two slits, as shown in Fig. 15–7. The diameter of the solenoid is to be much smaller than the distance d between the two slits. In these circumstances, the diffraction of the electrons at the slit gives no appreciable probability that the electrons will get near the solenoid. What will be the effect on our interference experiment?

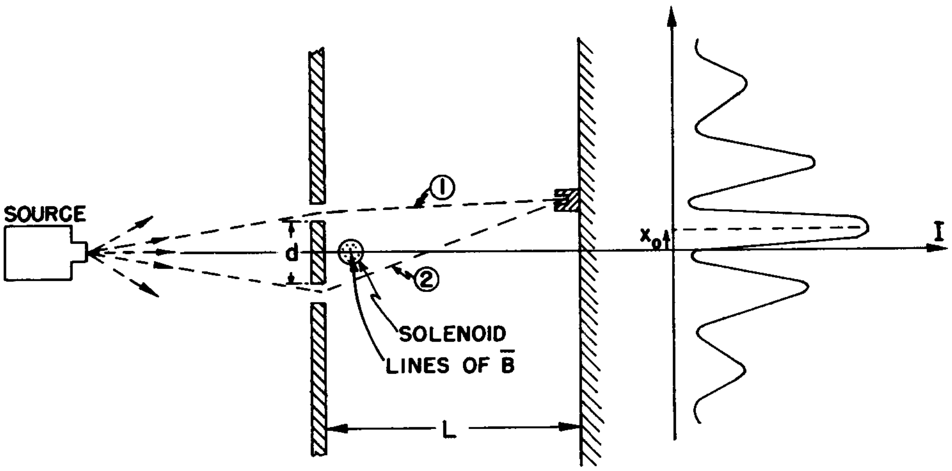


Fig. 15–7. A magnetic field can influence the motion of electrons even though it exists only in regions where there is an arbitrarily small probability of finding the electrons.

We compare the situation with and without a current through the solenoid. If we have no current, we have no $\boldsymbol{B}$ or $\boldsymbol{A}$ and we get the original pattern of electron intensity at the backstop. If we turn the current on in the solenoid and build up a magnetic field $\boldsymbol{B}$ inside, then there is an $\boldsymbol{A}$ outside. There is a shift in the phase difference proportional to the circulation of $\boldsymbol{A}$ outside the solenoid, which will mean that the pattern of maxima and minima is shifted to a new position. In fact, since the flux of $\boldsymbol{B}$ inside is a constant for any pair of paths, so also is the circulation of $\boldsymbol{A}$. For every arrival point there is the same phase change; this corresponds to shifting the entire pattern in x by a constant amount, say x_0, that we can easily calculate. The maximum intensity will occur where the phase difference between the two waves is zero. Using Eq. (15.32) or Eq. (15.33) for δ and Eq. (15.28) for $\delta(B = 0)$, we have

$$x_0 = -\frac{L}{d}\,\bar{\lambda}\,\frac{q}{\hbar}\oint_{(1-2)} \boldsymbol{A}\cdot d\boldsymbol{s}, \tag{15.35}$$

or

$$x_0 = -\frac{L}{d}\,\bar{\lambda}\,\frac{q}{\hbar}\,[\text{flux of } \boldsymbol{B} \text{ between (1) and (2)}]. \tag{15.36}$$

The pattern with the solenoid in place should appear* as shown in Fig. 15–7. At least, that is the prediction of quantum mechanics.

Precisely this experiment has recently been done. It is a very, very difficult experiment. Because the wavelength of the electrons is so small, the apparatus must be on a tiny scale to observe the interference. The slits must be very close together, and that means that one needs an exceedingly small solenoid. It turns out that in certain circumstances, iron crystals will grow in the form of very long, microscopically thin filaments called whiskers. When these iron whiskers are magnetized they are like a tiny solenoid, and there is no field outside except near the ends. The electron interference experiment was done with such a whisker between two slits, and the predicted displacement in the pattern of electrons was observed.

In our sense then, the $\boldsymbol{A}$-field is "real." You may say: "But there *was* a magnetic field." There was, but remember our original idea—that a field is "real" if it is what must be specified *at the position* of the particle in order to get the motion. The $\boldsymbol{B}$-field in the whisker acts at a distance. If we want to describe its influence not as action-at-a-distance, we must use the vector potential.

This subject has an interesting history. The theory we have described was known from the beginning of quantum mechanics in 1926. The fact that the vector potential appears in the wave equation of quantum mechanics (called the Schrödinger equation) was obvious from the day it was written. That it cannot be replaced by the magnetic field in any easy way was observed by one man after the other who tried to do so. This is also clear from our example of electrons moving in a region where there is no field and being affected nevertheless. But because in classical mechanics $\boldsymbol{A}$ did not appear to have any direct importance and, furthermore, because it could be changed by adding a gradient, people repeatedly said that the vector potential had no direct physical significance—that only the magnetic and electric fields are "right" even in quantum mechanics. It seems strange in retrospect that no one thought of discussing this experiment until 1956, when Bohm and Aharanov first suggested it and made the whole question crystal clear. The implication was there all the time, but no one paid attention to it. Thus many people were rather shocked when the matter was brought up. That's why someone thought it would be worth while to do the experiment to see that it really was right, even though quantum mechanics, which had been believed for so many years, gave an unequivocal answer. It is interesting that something like this can be around for thirty years but, because of certain prejudices of what is and is not significant, continues to be ignored.

Now we wish to continue in our analysis a little further. We will show the connection between the quantum-mechanical formula and the classical formula—to show why it turns out that if we look at things on a large enough scale it will look as though the particles are acted on by a force equal to $q\boldsymbol{v} \times$ the curl of $\boldsymbol{A}$. To get classical mechanics from quantum mechanics, we need to consider cases in which all the wavelengths are very small compared with distances over which external conditions, like fields, vary appreciably. We shall not prove the result in great generality, but only in a very simple example, to show how it works. Again we consider the same slit experiment. But instead of putting all the magnetic field in a very tiny region between the slits, we imagine a magnetic field that extends

* If the field $\boldsymbol{B}$ comes out of the plane of the figure, the flux as we have defined it is negative and x_0 is positive.

Gauge Transformations

ϕ determines E

and

A determines B

But ϕ and A are not unique

Gauge Transformations produce the different choices of ϕ and A that give the same E and B

Lots of Gauges

Coulomb Gauge
Lorenz Gauge
Axial Gauge
Temporal Gauge
Velocity Gauge
Kirchhoff Gauge
Landau Gauge
Feynman Gauge
t' Hooft Gauge
Unitary Gauge

B. Examples of gauges

The gauge invariance of classical field theory and of electrodynamics in particular allows one to consider the potential A^μ with various gauge conditions, most of them being not Poincaré invariant:

$$\partial_\mu A^\mu = 0 \quad (\mu = 0,1,23), \quad \text{Lorenz gauge,} \tag{34}$$

$$\nabla \cdot \mathbf{A} = \partial_j A_j = 0 \quad (j = 1,2,3),$$

$$\text{Coulomb gauge or radiation gauge,} \tag{35}$$

$$n_\mu A^\mu = 0 \quad (n^2 = 0), \qquad \text{light-cone gauge,} \tag{36}$$

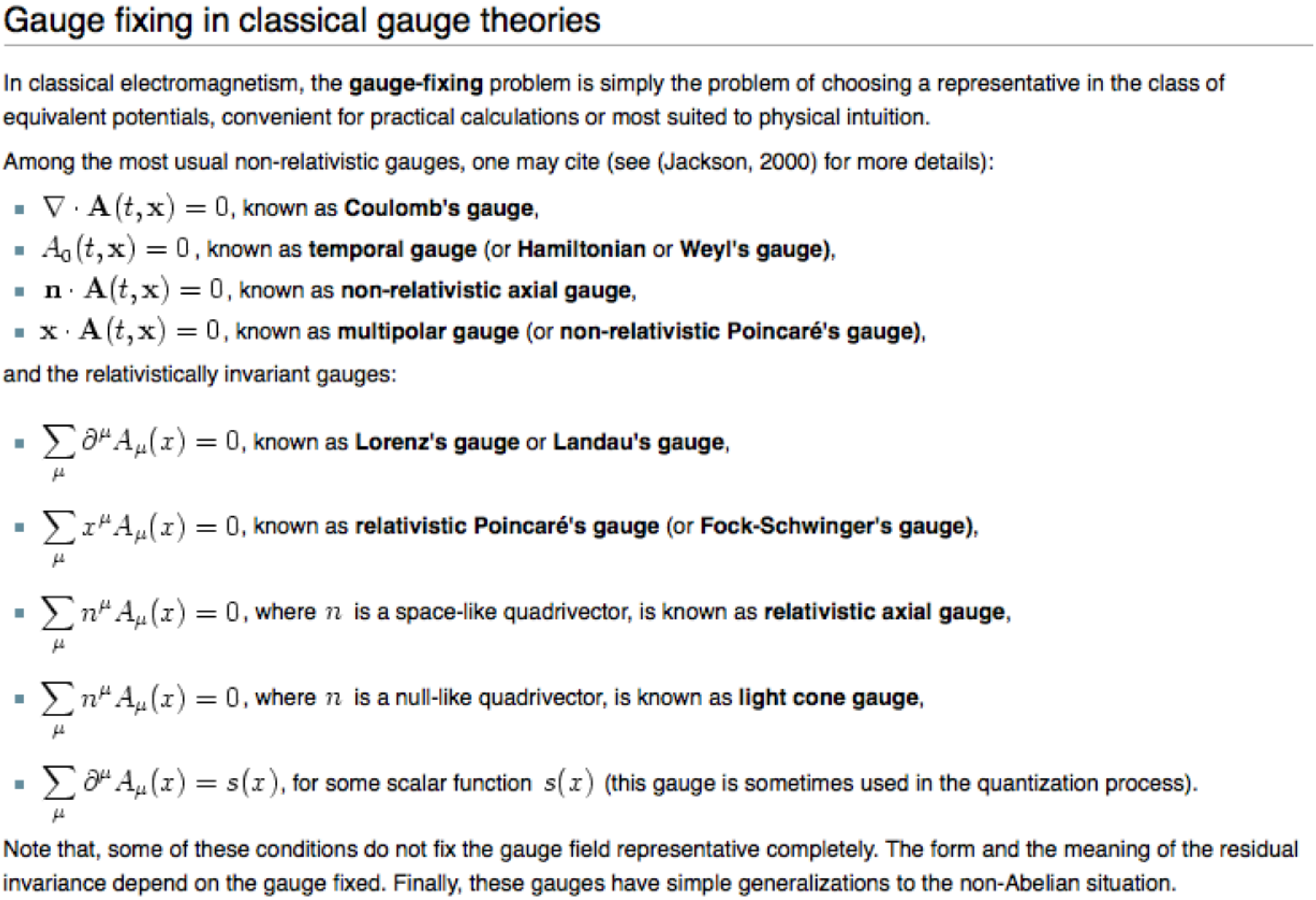
$$A_0 = 0, \qquad \text{Hamiltonian or temporal gauge,} \tag{37}$$

$$A_3 = 0, \qquad \text{axial gauge,} \tag{38}$$

$$x_\mu A^\mu = 0, \qquad \text{Fock-Schwinger gauge,} \tag{39}$$

$$x_j A_j = 0, \qquad \text{Poincaré gauge.} \tag{40}$$

An appropriate choice of gauge simplifies calculations.



Gauge fixing in classical gauge theories

In classical electromagnetism, the **gauge-fixing** problem is simply the problem of choosing a representative in the class of equivalent potentials, convenient for practical calculations or most suited to physical intuition.

Among the most usual non-relativistic gauges, one may cite (see (Jackson, 2000) for more details):

- $\nabla \cdot \mathbf{A}(t,\mathbf{x}) = 0$, known as **Coulomb's gauge**,
- $A_0(t,\mathbf{x}) = 0$, known as **temporal gauge** (or **Hamiltonian** or **Weyl's gauge**),
- $\mathbf{n} \cdot \mathbf{A}(t,\mathbf{x}) = 0$, known as **non-relativistic axial gauge**,
- $\mathbf{x} \cdot \mathbf{A}(t,\mathbf{x}) = 0$, known as **multipolar gauge** (or **non-relativistic Poincaré's gauge**),

and the relativistically invariant gauges:

- $\sum_\mu \partial^\mu A_\mu(x) = 0$, known as **Lorenz's gauge** or **Landau's gauge**,
- $\sum_\mu x^\mu A_\mu(x) = 0$, known as **relativistic Poincaré's gauge** (or **Fock-Schwinger's gauge**),
- $\sum_\mu n^\mu A_\mu(x) = 0$, where n is a space-like quadrivector, is known as **relativistic axial gauge**,
- $\sum_\mu n^\mu A_\mu(x) = 0$, where n is a null-like quadrivector, is known as **light cone gauge**,
- $\sum_\mu \partial^\mu A_\mu(x) = s(x)$, for some scalar function $s(x)$ (this gauge is sometimes used in the quantization process).

Note that, some of these conditions do not fix the gauge field representative completely. The form and the meaning of the residual invariance depend on the gauge fixed. Finally, these gauges have simple generalizations to the non-Abelian situation.

Historical roots of gauge invariance

J. D. Jackson*

University of California and Lawrence Berkeley National Laboratory, Berkeley, California 94720

L. B. Okun†

Institute of Theoretical and Experimental Physics, State Science Center of Russian Federation, 117218 Moscow, Russia

(Published 14 September 2001)

Gauge invariance is the basis of the modern theory of electroweak and strong interactions (the so-called standard model). A number of authors have discussed the ideas and history of quantum guage theories, beginning with the 1920s, but the roots of gauge invariance go back to the year 1820 when electromagnetism was discovered and the first electrodynamic theory was proposed. We describe the 19th century developments that led to the discovery that different forms of the vector potential (differing by the gradient of a scalar function) are physically equivalent, if accompanied by a change in the scalar potential: $\mathbf{A}\to\mathbf{A}'=\mathbf{A}+\nabla\chi$, $\Phi\to\Phi'=\Phi-\partial\chi/c\partial t$. L. V. Lorenz proposed the condition $\partial_\mu A^\mu=0$ in the mid-1860s, but this constraint is generally misattributed to the better known H. A. Lorentz. In the work in 1926 on the relativistic wave equation for a charged spinless particle in an electromagnetic field by Schrödinger, Klein, and Fock, it was Fock who discovered the invariance of the equation under the above changes in $\mathbf{A}$ and Φ if the wave function was transformed according to $\psi\to\psi'=\psi\exp(ie\chi/\hbar c)$. In 1929, H. Weyl proclaimed this invariance as a general principle and called it *Eichinvarianz* in German and *gauge invariance* in English. The present era of non-Abelian gauge theories started in 1954 with the paper by Yang and Mills on isospin gauge invariance.

CONTENTS

I. INTRODUCTION

The principle of gauge invariance plays a key role in the standard model, which describes electroweak and strong interactions of elementary particles. Its origins can be traced to Fock (1926b), who extended the known freedom of choosing the electromagnetic potentials in classical electrodynamics to the quantum mechanics of charged particles interacting with electromagnetic fields. Equations (5) and (9) of Fock's paper are, in his notation,

$$\mathbf{A}=\mathbf{A}_1+\nabla f,$$

$$\varphi=\varphi_1-\frac{1}{c}\frac{\partial f}{\partial t}, \qquad \text{Fock's (5)}$$

$$p=p_1-\frac{e}{c}f,$$

and

$$\psi=\psi_0 e^{2\pi i p/h}. \qquad \text{Fock's (9)}$$

In present-day notation we write

$$\mathbf{A}\to\mathbf{A}'=\mathbf{A}+\nabla\chi, \tag{1a}$$

$$\Phi\to\Phi'=\Phi-\frac{1}{c}\frac{\partial\chi}{\partial t}, \tag{1b}$$

$$\psi\to\psi'=\psi\exp(ie\chi/\hbar c). \tag{1c}$$

Here $\mathbf{A}$ is the vector potential, Φ is the scalar potential, and χ is known as the gauge function. The Maxwell equations of classical electromagnetism for the electric and magnetic fields are invariant under the transformations (1a) and (1b) of the potentials. What Fock discovered was that, for the quantum dynamics, that is, the form of the quantum equation, to remain unchanged by these transformations, the wave function must undergo the transformation (1c), whereby it is multiplied by a local (space-time-dependent) phase. The concept was declared a general principle and "consecrated" by Hermann Weyl (1928, 1929a, 1929b). The invariance of a

*Electronic address: jdj@lbl.gov
†Electronic address: okun@heron.itep.ru

0034-6861/2001/73(3)/663(18)/$23.60

$$\mathbf{A} \rightarrow \mathbf{A}' = \mathbf{A} + \boldsymbol{\nabla} \chi,$$

$$\Phi \rightarrow \Phi' = \Phi - \frac{1}{c} \frac{\partial \chi}{\partial t},$$

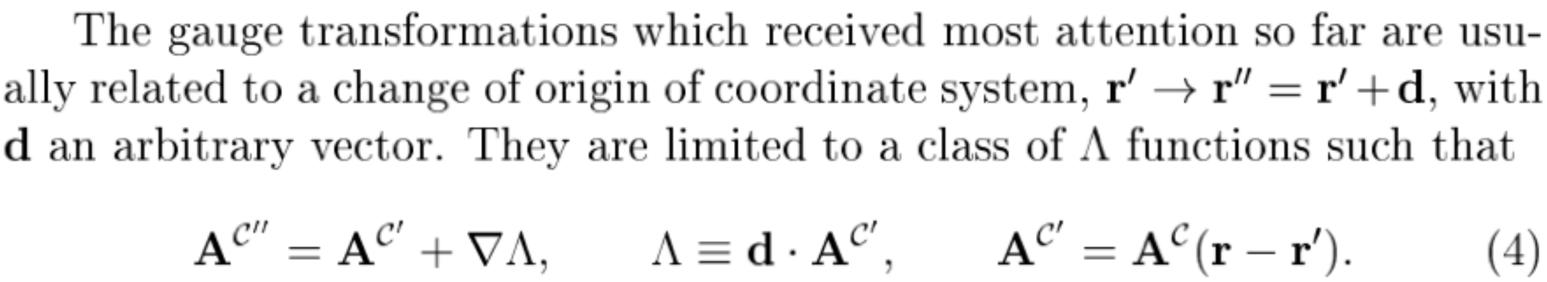
$$\psi \rightarrow \psi' = \psi \exp(ie\chi/\hbar c).$$

Simplest EM gauge transformations

Scalar potential

V => V' = V + c

Vector potential



The gauge transformations which received most attention so far are usually related to a change of origin of coordinate system, $\mathbf{r}' \to \mathbf{r}'' = \mathbf{r}' + \mathbf{d}$, with $\mathbf{d}$ an arbitrary vector. They are limited to a class of Λ functions such that

$$\mathbf{A}^{\mathcal{C}''} = \mathbf{A}^{\mathcal{C}'} + \nabla\Lambda, \qquad \Lambda \equiv \mathbf{d} \cdot \mathbf{A}^{\mathcal{C}'}, \qquad \mathbf{A}^{\mathcal{C}'} = \mathbf{A}^{\mathcal{C}}(\mathbf{r} - \mathbf{r}'). \tag{4}$$

The flux of the external magnetic field $\mathbf{B} = \nabla \times \mathbf{A}^{\mathcal{C}}$ is obviously invariant in a gauge transformation [8]

$$\mathbf{A}^{\mathcal{C}} \rightarrow \mathbf{A}^{\mathcal{C}} + \nabla\Lambda, \tag{2}$$

where $\Lambda(\mathbf{r})$ is an arbitary function of position, well behaved for $r \rightarrow \infty$. When a gauge transformation is carried out, the wave function ψ and the Hamiltonian h of a particle with charge q are transformed according to the equations

$$\psi \rightarrow \psi' = \psi \exp\left(\frac{iq}{\hbar c}\Lambda\right) \qquad h \rightarrow h' = \exp\left(\frac{iq}{\hbar c}\Lambda\right) h \exp\left(-\frac{iq}{\hbar c}\Lambda\right), \tag{3}$$

leaving invariant the one particle Schrödinger equation $h\psi = \epsilon\psi$. If the basis set is complete, also the calculated energy of the system is gauge-invariant, and the electronic current density is conserved, i.e., $\nabla \cdot \mathbf{j} = 0$ [2]. For wavefunctions that are exact eigenfunctions to some model hamiltonian, the choice of gauge is immaterial, because variations of the diamagnetic and paramagnetic contributions arising from a change of gauge cancel out one another, leaving invariant total magnetic properties. More generally, in the case of actual calculations relying on the algebraic approximation, selecting a gauge, e.g., specifying the "best" form of the vector potential, introduces serious problems: the fulfillment of constraints for gauge invariance constitutes a severe test of reliability for the computational scheme, and a hallmark of quality of the molecular wave function.

The gauge transformations which received most attention so far are usually related to a change of origin of coordinate system, $\mathbf{r}' \rightarrow \mathbf{r}'' = \mathbf{r}' + \mathbf{d}$, with $\mathbf{d}$ an arbitrary vector. They are limited to a class of Λ functions such that

$$\mathbf{A}^{\mathcal{C}''} = \mathbf{A}^{\mathcal{C}'} + \nabla\Lambda, \qquad \Lambda \equiv \mathbf{d} \cdot \mathbf{A}^{\mathcal{C}'}, \qquad \mathbf{A}^{\mathcal{C}'} = \mathbf{A}^{\mathcal{C}}(\mathbf{r} - \mathbf{r}'). \tag{4}$$

More general gauge transformations have been also proposed. The Landau gauge [8] for computing magnetic properties was studied in a series of articles [9]-[12].

Within Coulomb and Landau gauges, the vector potential is purely transverse; an alternative form of vector potential containing a longitudinal component has also been considered [13]. Some interesting features of these gauges were pointed out elsewhere [9]- [13].

the term more difficult to evaluate accurately, i.e., the paramagnetic contribution to magnetic susceptibility, by means of a gauge transformation. It is shown that an absolute minimum in a variational sense cannot be determined a priori. However, a 'local' minimum of the paramagnetic contribution to magnetic susceptibility can be arrived at by employing general gauge transformations of polynomial form.

I Introduction

The gauge problem affecting theoretical evaluation of response properties of a molecule in the presence of an external magnetic field arises from the unphysical dependence of calculated values on the non-unique definition of the vector potential.

It is well-established, from theoretical [1, 2], as well as computational point of view [3, 4], that coupled Hartree-Fock (CHF) methods [1] yield magnetic susceptibility and nuclear magnetic shielding invariant to a gauge transformation only in the limit of complete atomic basis sets. This requirement brings in serious difficulties, for, in most cases, use of large basis sets is mandatory to achieve near Hartree-Fock accuracy, and, at the same time, a satisfactory degree of gauge independence.

Several computational procedures have been proposed so far to circumvent the problem of gauge translation, and related invariance of calculated magnetic properties in a change of coordinate system. Current methods exploit atomic basis sets explicitly depending on a gauge factor (GIAO) [5], individual gauges for localized molecular orbitals (IGLO) [6], and (LORG) [7].

Theoretical determinations of magnetic properties of a molecule perturbed by an external magnetic field, assumed, for the sake of simplicity, spatially uniform and time independent, are usually obtained by retaining the Coulomb gauge for the vector potential $\mathbf{A}^{\mathcal{C}}$ [8]:

$$\mathbf{A}^{\mathcal{C}}(\mathbf{r}) = \frac{1}{2}\mathbf{B} \times \mathbf{r}, \qquad \nabla \cdot \mathbf{A}^{\mathcal{C}} = 0. \tag{1}$$

[†] To whom correspondence should be addressed.

[‡] Members of Carrera del Investigador del CONICET.

General Relativity

The gauge transformations for general relativity are arbitrary coordinate transformations. Technically, the transformations must be invertible, and both the transformation and its inverse must be smooth, in the sense of being differentiable an arbitrary number of times.

Einstein's Field Equations

In Einstein's general relativity, coordinates like x, y, z, and t are not only "relative" in the global sense of translations like t \rightarrow t+C, rotations, etc., but become completely arbitrary, so that for example one can define an entirely new timelike coordinate according to some arbitrary rule such as

$$t \Rightarrow t' = t^2/t_o$$

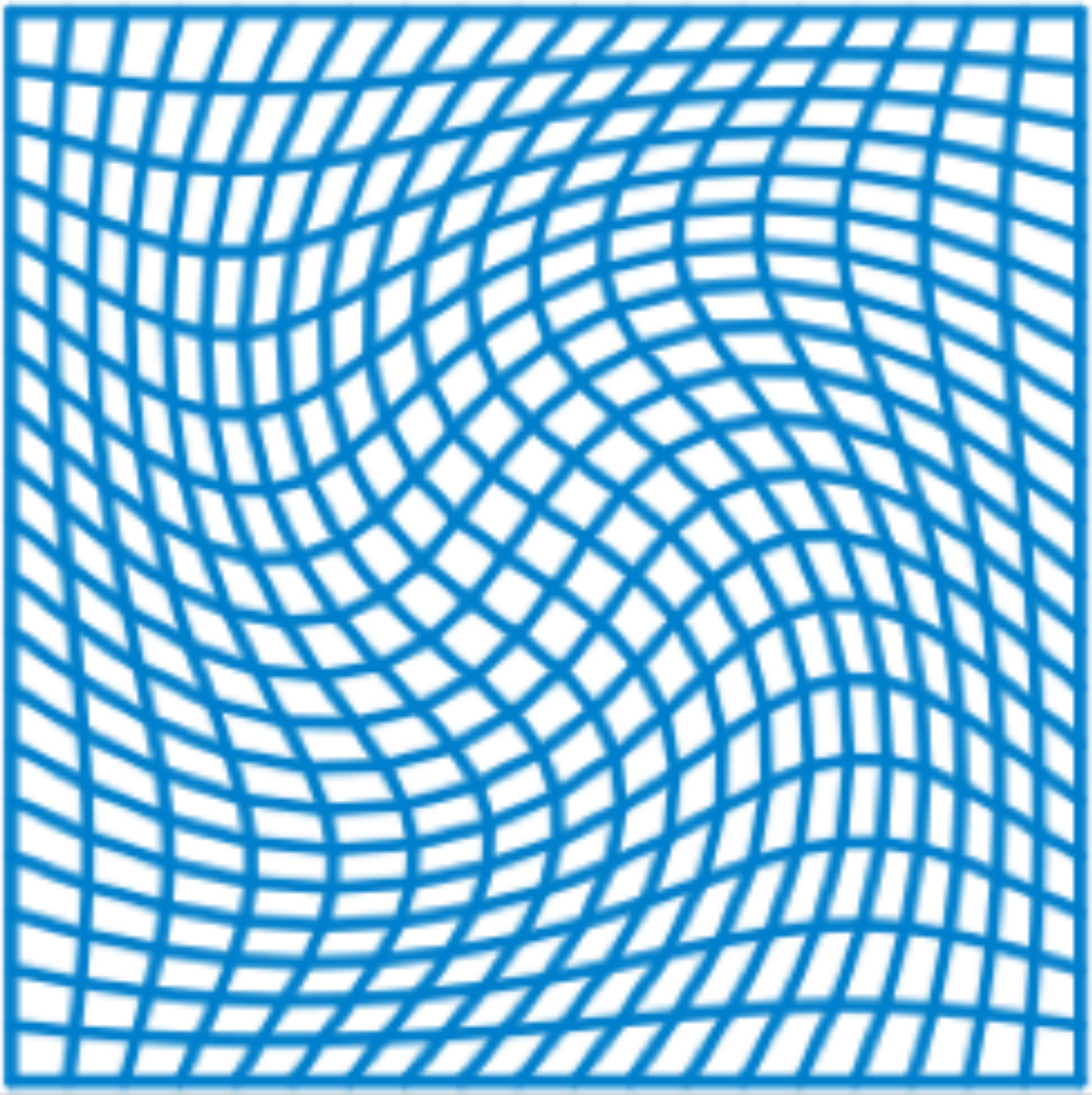
where t_o has units of time, and yet Einstein's equations will have the same form.

Invariance of the form of an equation under an arbitrary coordinate transformation is customarily referred to as general covariance and equations with this property are referred to as written in the covariant form. General covariance is a subclass of gauge invariance.

Maxwell's equations can also be expressed in a generally covariant form, which is as invariant under general coordinate transformation as are Einstein's Field Equations.

Translation Invariance

Space translation $r \Rightarrow r' = r + 1$ light year
Time translation $t \Rightarrow t' = t + 10$ million years
These are global symmetries



The Cartesian coordinate grid on this square has been distorted by a coordinate transformation, so that there is a nonlinear relationship between the old (x,y) coordinates and the new ones. Einstein's equations of general relativity are still valid in the new coordinate system. Such changes of coordinate system are the gauge transformations of general relativity.

TABLE 3.1 ANALOGY BETWEEN THE ELECTROMAGNETIC AND GRAVITATIONAL FIELD THEORIES

	Electromagnetism	**Gravitation (Linear Approximation)**
Source of field	j^ν	$T^{\mu\nu}$
Conservation law	$\partial_\nu j^\nu = 0$	$\partial_\nu T^{\mu\nu} = 0$
Field	A^ν	$h^{\mu\nu}$
Field equation	$\partial_\mu \partial^\mu A^\nu - \partial^\nu \partial_\mu A^\mu = 4\pi j^\nu$	$\partial_\lambda \partial^\lambda h^{\mu\nu} - 2\partial_\lambda \partial^{(\nu} h^{\mu)\lambda} + \partial^\mu \partial^\nu h$ $-\eta^{\mu\nu}\partial_\lambda \partial^\lambda h + \eta^{\mu\nu}\partial_\lambda \partial_\sigma h^{\lambda\sigma} = -\kappa T^{\mu\nu}$
Gauge transformation	$A^\mu \rightarrow A^\mu + \partial^\mu \Lambda$	$h^{\mu\nu} \rightarrow h^{\mu\nu} + \partial^{(\nu}\Lambda^{\mu)}$
Preferred gauge condition	$\partial_\mu A^\mu = 0$	$\partial_\mu(h^{\mu\nu} - \frac{1}{2}\eta^{\mu\nu}h) = 0$
Field equation in preferred gauge	$\partial_\mu \partial^\mu A^\nu = 4\pi j^\nu$	$\partial_\lambda \partial^\lambda (h^{\mu\nu} - \frac{1}{2}\eta^{\mu\nu}h) = -\kappa T^{\mu\nu}$
Energy-momentum exchange between field and particle	$\partial_\nu T_{(m)\mu}{}^\nu = F_{\mu\nu} j^\nu$	$\partial_\nu T_{(m)\mu}{}^\nu = \dfrac{\kappa}{2} m h_{\alpha\beta,\mu} T_{(m)}{}^{\alpha\beta}$
Equation of motion of particle	$\dfrac{d}{d\tau} p_\mu = q F_{\mu\nu} u^\nu$	$\dfrac{d}{d\tau} P_\mu = \dfrac{\kappa}{2} m h_{\alpha\beta,\mu} u^\alpha u^\beta$
Energy-momentum of particle	$p_\mu = m u_\mu$	$P_\mu = m u_\mu + m\kappa h_{\mu\alpha} u^\alpha$
Proper time interval	$d\tau^2 = \eta_{\alpha\beta} dx^\alpha dx^\beta$	$d\tau^2 = (\eta_{\alpha\beta} + \kappa h_{\alpha\beta}) dx^\alpha dx^\beta$

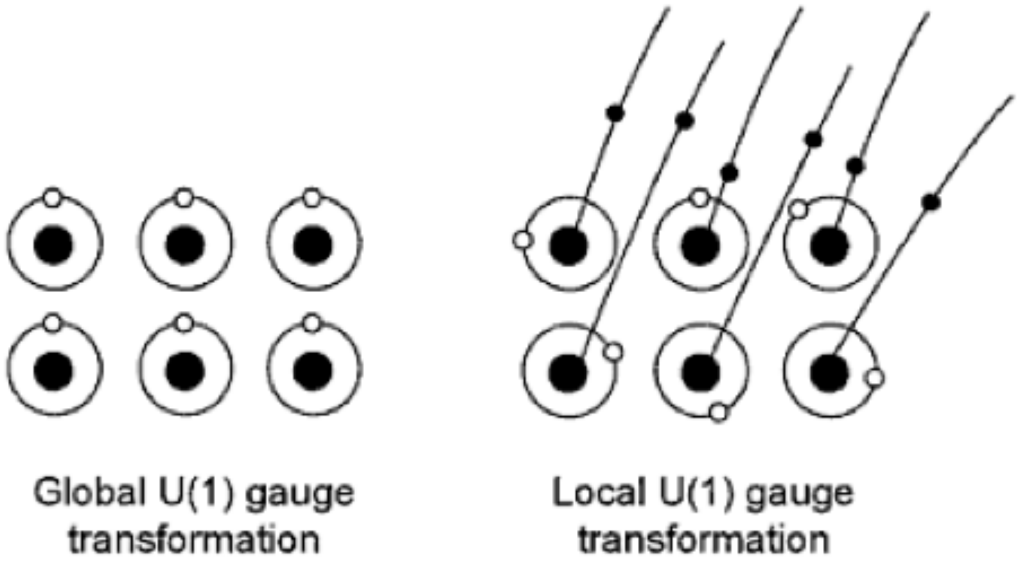

Global U(1) gauge transformation
Local U(1) gauge transformation

Fig. 7.4 Global and local gauge transformations. Left panel: in a global gauge transformation, the quantum phase change is the same at all space-time points. Right panel: in a local gauge transformation, the quantum phase can have arbitrary independent values at different space-time points, but it is correlated with the gauge field, whose gauge function (position on its fiber) keeps track of the phase.

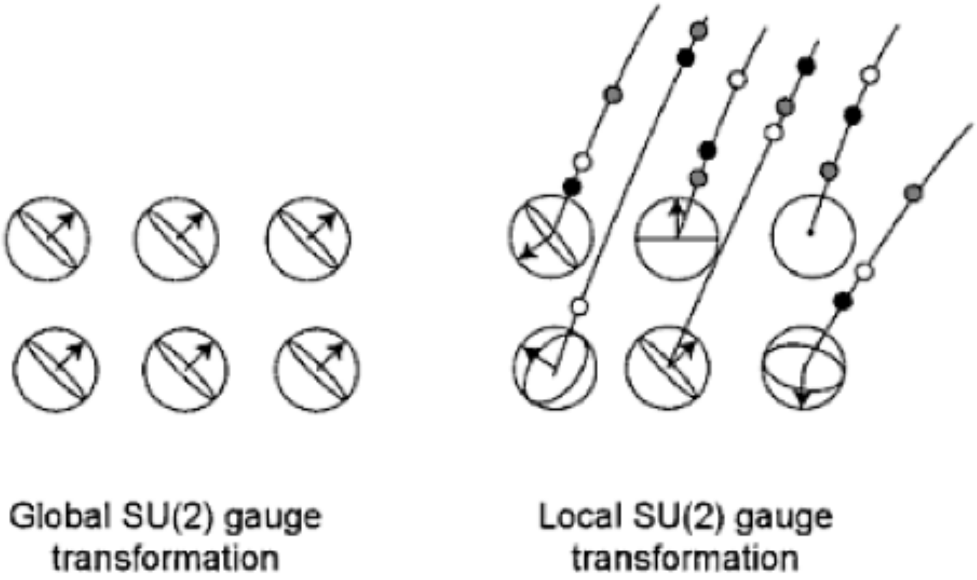

Global SU(2) gauge transformation
Local SU(2) gauge transformation

Fig. 9.3 In a global SU(2) gauge transformation, symbolic gyroscopes attached to points of space-time rotate in unision. In a local transformation, they rotate independently, but three gauge fields undergo correlated gauge transformations. The latter is indicated by the positions of three beads on a fiber. Compare with Fig. 7.4.

Gauge theories in elementary particle physics

This was the topic of the 1999 Nobel Prize. An idea was proposed by C.N. Yang and Robert Mills in 1954: they suggested that particles in the sub-atomic world might interact via fields that are similar to, but more general than electricity and magnetism. But, even though the interactions that had been registered in experiments showed some vague resemblance to the Yang-Mills equations, the details seemed to be all wrong. Attempts to perform accurate calculations were frustrated by infinite---hence meaningless---results. Together with my advisor then, and my co-Nobel-laureate now, M. Veltman, I found in 1970 how to renormalize the theory, and, more importantly, we identified the theories for which this works, and what conditions they must fullfil. One must, for instance, have a so-called Higgs-particle. These theories are now called gauge theories*.*

It was subsequently discovered that, indeed, the observed details of all known forces exactly agree with this picture. First it was found that the so-called weak force, in combination with the more familiar electro-magnetic one, is exactly described by a Yang-Mills theory. In 1973 it was concluded that also the strong force is a Yang-Mills theory. I was among the small number of people who were already convinced of this from early 1971. During the later 1970s, all pieces fell into place. Of all simple models describing the fundamental particles, one was standing out, the so-called Standard Model.

Gauge theories are the backbone of this Standard Model. But now it also became clear that this is much more than just a model: it is the Standard Theory. Great precision can be reached, though the practical difficulties in some sectors are still substantial, and it would be great if one could devise more powerful calculation techniques. Also, in spite of all its successes, the Standard Model, as it is formulated at present, shows deficiencies. It cannot be exactly right. Significant refinements are expected when the results of new experiments become known, hopefully during 2010 and subsequent years, when the European particle accelerator LHC becomes fully operational.

http://www.staff.science.uu.nl/~hooft101/

Will the Higgs be found?

More and more frequently, I receive letters and mails from wise people outside physics, telling me that "they know" that the Higgs will not be found, that our theories are baloney, how dare we spend billions of public funds to build machines such as LHC, "to prove, against better judgment, that our theories still stand a chance of being correct", and so on.

Well, dear friends, I am not going to answer all of you in person. Please do consider the scientific facts concerning the Standard Model. Fact is that the W+, W? and the Z boson each carry three spin degrees of freedom, whereas the Yang-Mills field quanta, which describe their interactions correctly in great detail, each carry only two. Those remaining modes come from the Higgs field. What this means is that three quarters of the field of the Higgs have already been found. The fourth is still missing, and if you calculate its properties, it is also clear why it is missing: it is hiding in the form of a particle that is difficult to detect. LHC will have to work for several years before it stands a chance to see the statistical signals of this Higgs particle. What compounds the matter even more is that there may well be several sets of Higgs fields. If there are two, which is eight quarters of the field, we will get five Higgses rather than one. This would be a quite realistic possibility but it would make the detection of each one of them even harder, because they cause more complex statistical signals that are more difficult to predict.

Theories without any Higgs particle are possible but ugly and have been practically ruled out by observations. In such theories, composite bound states of other particles have to play the role of a Higgs, which requires the existence of very strong new interactions, of which there is presently no evidence at all, and it would make the perfect agreement found today between observations and the Standard Model highly improbable.

If no Higgs is found at all, and all present ideas would be ruled out by LHC - which is unlikely - this would in no way make LHC useless. Quite to the contrary, this would lead to lots of work for theoreticians to do, and, more importantly, this would imply the existence of new strong forces with a plethora of highly interesting particles just round the corner, waiting to be discovered by LHC or its successors. There's no way that this field of research can become dull.

http://www.staff.science.uu.nl/~hooft101/

What is a gauge?

27 September, 2008 in expository, math.AP, math.CO, math.DG, math.DS, math.MP | Tags: connections, curvature, fibre bundles, gauge fixing, gauge invariance, sections | by Terence Tao

"Gauge theory" is a term which has connotations of being a fearsomely complicated part of mathematics – for instance, playing an important role in quantum field theory, general relativity, geometric PDE, and so forth. But the underlying concept is really quite simple: a *gauge* is nothing more than a "coordinate system" that varies depending on one's "location" with respect to some "base space" or "parameter space", a *gauge transform* is a change of coordinates applied to each such location, and a *gauge theory* is a model for some physical or mathematical system to which gauge transforms can be applied (and is typically *gauge invariant*, in that all physically meaningful quantities are left unchanged (or transform naturally) under gauge transformations). By *fixing* a gauge (thus *breaking* or *spending* the gauge symmetry), the model becomes something easier to analyse mathematically, such as a system of partial differential equations (in classical gauge theories) or a perturbative quantum field theory (in quantum gauge theories), though the tractability of the resulting problem can be heavily dependent on the choice of gauge that one fixed. Deciding exactly how to fix a gauge (or whether one should spend the gauge symmetry at all) is a key question in the analysis of gauge theories, and one that often requires the input of geometric ideas and intuition into that analysis.

I was asked recently to explain what a gauge theory was, and so I will try to do so in this post. For simplicity, I will focus exclusively on classical gauge theories; quantum gauge theories are the quantization of classical gauge theories and have their own set of conceptual difficulties (coming from quantum field theory) that I will not discuss here. While gauge theories originated from physics, I will not discuss the physical significance of these theories much here, instead focusing just on their mathematical aspects. My discussion will be informal, as I want to try to convey the geometric intuition rather than the rigorous formalism (which can, of course, be found in any graduate text on differential geometry).

– Coordinate systems –

Before I discuss gauges, I first review the more familiar concept of a *coordinate system*, which is basically the special case of a gauge when the base space (or parameter space) is trivial.

Classical mathematics, such as practised by the ancient Greeks, could be loosely divided into two disciplines, *geometry* and *number theory*, where I use the latter term very broadly, to encompass all sorts of mathematics dealing with any sort of number. The two disciplines are unified by the concept of a *coordinate system*, which allows one to convert geometric objects to numeric ones or vice versa. The most well known example of a coordinate system is the Cartesian coordinate system for the plane (or more generally for a Euclidean space), but this is just one example of many such systems. For instance:

1. One can convert a length (of, say, an interval) into an (unsigned) real number, or vice versa, once one fixes a unit of length (e.g. the metre or the foot). In this case, the coordinate system is specified by the choice of length unit.
2. One can convert a displacement along a line into a (signed) real number, or vice versa, once one

fixes a unit of length *and* an orientation along that line. In this case, the coordinate system is specified by the length unit together with the choice of orientation. Alternatively, one can replace the unit of length and the orientation by a unit displacement vector e along the line.

3. One can convert a position (i.e. a point) on a line into a real number, or vice versa, once one fixes a unit of length, an orientation along the line, *and* an origin on that line. Equivalently, one can pick an origin O and a unit displacement vector e. This coordinate system essentially identifies the original line with the standard real line $\mathbb{R}$.
4. One can generalise these systems to higher dimensions. For instance, one can convert a displacement along a plane into a vector in $\mathbb{R}^2$, or vice versa, once one fixes two linearly independent displacement vectors e_1, e_2 (i.e. a basis) to span that plane; the Cartesian coordinate system is just one special case of this general scheme. Similarly, one can convert a position on a plane to a vector in $\mathbb{R}^2$ once one picks a basis e_1, e_2 for that plane as well as an origin O, thus identifying that plane with the standard Euclidean plane $\mathbb{R}^2$. (To put it another way, units of measurement are nothing more than one-dimensional (i.e. scalar) coordinate systems.)
5. To convert an angle in a plane to a signed number (modulo multiples of 2π), or vice versa, one needs to pick an orientation on the plane (e.g. to decide that anti-clockwise angles are positive).
6. To convert a *direction* in a plane to a signed number (again modulo multiples of 2π), or vice versa, one needs to pick an orientation on the plane, as well as a reference direction (e.g. true or magnetic north is often used in the case of ocean navigation).
7. Similarly, to convert a position on a circle to a number (modulo multiples of 2π), or vice versa, one needs to pick an orientation on that circle, together with an origin on that circle. Such a coordinate system then equates the original circle to the standard unit circle $S^1 := \{z \in \mathbb{C} : |z| = 1\}$ (with the standard origin $+1$ and the standard anticlockwise orientation $\circlearrowleft$).
8. To convert a position on a two-dimensional sphere (e.g. the surface of the Earth, as a first approximation) to a point on the standard unit sphere $S^2 := \{(x,y,z) \in \mathbb{R}^3 : x^2 + y^2 + z^2\}$, one can pick an orientation on that sphere, an "origin" (or "north pole") for that sphere, and a "prime meridian" connecting the north pole to its antipode. Alternatively, one can view this coordinate system as determining a pair of Euler angles ϕ, λ (or a latitude and longitude) to be assigned to every point on one's original sphere.
9. The above examples were all geometric in nature, but one can also consider "combinatorial" coordinate systems, which allow one to identify combinatorial objects with numerical ones. An extremely familiar example of this is *enumeration*: one can identify a set A of (say) five elements with the numbers 1,2,3,4,5 simply by choosing an enumeration $a_1, a_2, \ldots, a_5$ of the set A. One can similarly enumerate other combinatorial objects (e.g. graphs, relations, trees, partial orders, etc.), and indeed this is done all the time in combinatorics. Similarly for algebraic objects, such as cosets of a subgroup H (or more generally, torsors of a group G); one can identify such a coset with H itself by designating an element of that coset to be the "identity" or "origin".

More generally, a coordinate system Φ can be viewed as an isomorphism $\Phi: A \to G$ between a given geometric (or combinatorial) object A in some class (e.g. a circle), and a standard object G in that class (e.g. the standard unit circle). (To be pedantic, this is what a *global* coordinate system is; a *local* coordinate system, such as the coordinate charts on a manifold, is an isomorphism between a local piece of a geometric or combinatorial object in a class, and a local piece of a standard object in that class. I will restrict attention to global coordinate systems for this discussion.)

Coordinate systems identify geometric or combinatorial objects with numerical (or standard) ones, but in many cases, there is no natural (or canonical) choice of this identification; instead, one may be faced with a variety of coordinate systems, all equally valid. One can of course just fix one such system once and for all, in which case there is no real harm in thinking of the geometric and numeric objects as being equivalent. If however one plans to change from one system to the next (or to avoid using such systems altogether), then it becomes important to carefully distinguish these two types of objects, to avoid confusion. For instance, if an interval AB is measured to have a length of 3 yards, then it is OK to write $|AB| = 3$ (identifying the geometric concept of length with the numeric concept of a positive real number) so long as you plan to stick to having the yard as the unit of length for the rest of one's analysis. But if one was also planning to use, say, feet, as a unit of length also, then to avoid confusing statements such as "$|AB| = 3$ and $|AB| = 9$", one should specify the coordinate systems explicitly, e.g. "$|AB| = 3 \hbox{ yards}$ and $|AB| = 9 \hbox{ feet}$". Similarly, identifying a point P in a plane with its coordinates (e.g. $P = (4,3)$) is safe as long as one intends to only use a single coordinate system throughout; but if one intends to change coordinates at some point (or to switch to a coordinate-free perspective) then one should be more careful, e.g. writing $P = 4e_1 + 3e_2$, or even $P = O + 4e_1 + 3e_2$, if the origin O and basis vectors e_1, e_2 of one's coordinate systems might be subject to future change.

As mentioned above, it is possible to in many cases to dispense with coordinates altogether. For instance, one can view the length $|AB|$ of a line segment AB not as a number (which requires one to select a unit of length), but more abstractly as the equivalence class of all line segments CD that are congruent to AB. With this perspective, $|AB|$ no longer lies in the standard semigroup $\mathbb{R}^+$, but in a more abstract semigroup $\mathcal{L}$ (the space of line segments quotiented by congruence), with addition now defined geometrically (by concatenation of intervals) rather than numerically. A unit of length can now be viewed as just one of many different isomorphisms $\Phi: \mathcal{L} \to \mathbb{R}^+$ between $\mathcal{L}$ and $\mathbb{R}^+$, but one can abandon the use of such units and just work with $\mathcal{L}$ directly. Many statements in Euclidean geometry involving length can be phrased in this manner. For instance, if B lies in AC, then the statement $|AC| = |AB| + |BC|$ can be stated in $\mathcal{L}$, and does not require any units to convert $\mathcal{L}$ to $\mathcal{R}^+$; with a bit more work, one can also make sense of such statements as $|AC|^2 = |AB|^2 + |BC|^2$ for a right-angled triangle ABC (i.e. Pythagoras' theorem) while avoiding units, by defining a symmetric bilinear product operation $\times: \mathcal{L} \times \mathcal{L} \to \mathcal{A}$ from the abstract semigroup $\mathcal{L}$ of lengths to the abstract semigroup $\mathcal{A}$ of areas. (Indeed, this is basically how the ancient Greeks, who did not quite possess the modern real number system $\mathbb{R}$, viewed geometry, though of course without the assistance of such modern terminology as "semigroup" or "bilinear".)

The above abstract *coordinate-free perspective* is equivalent to a more concrete *coordinate-invariant perspective*, in which we do allow the use of coordinates to convert all geometric quantities to numeric ones, but insist that every statement that we write down is invariant under changes of coordinates. For instance, if we shrink our chosen unit of length by a factor $\lambda > 0$, then the numerical length of every interval increases by a factor of λ, e.g. $|AB| \mapsto \lambda |AB|$. The coordinate-invariant approach to length measurement then treats lengths such as $|AB|$ as numbers, but requires all statements involving such lengths to be invariant under the above scaling symmetry. For instance, a statement such as $|AC|^2 = |AB|^2 + |BC|^2$ is legitimate under this perspective, but a statement such as $|AB| = |BC|^2$ or $|AB| = 3$ is not. [In other words, co-ordinate invariance here is the same thing as being dimensionally consistent. Indeed, dimensional analysis is nothing more than the analysis of the scaling symmetries in one's coordinate systems.] One can retain this coordinate-invariance symmetry throughout one's arguments; or one can, at some point, choose to *spend* (or *break*) this coordinate invariance by selecting (or *fixing*) the coordinate system (which, in this case, means selecting a unit length). The advantage in spending such a symmetry is that one can often normalise one or more quantities to equal a particularly nice value; for instance, if a length $|AB|$ is appearing everywhere in one's arguments, and one has carefully retained coordinate-invariance up until some key point, then it can be convenient to spend this invariance to normalise $|AB|$ to equal 1. (In this case, one only has a one-dimensional family of symmetries, and so can only normalise one quantity at a time; but when one's symmetry group is larger, one can often normalise many more quantities at once; as a rule of thumb, one can normalise one quantity for each degree of freedom in the symmetry group.) Conversely, if one has already spent the coordinate invariance, one can often buy it back by converting all the facts, hypotheses, and desired conclusions one currently possesses in the situation back to a coordinate-invariant formulation. Thus one could imagine performing one normalisation to do one set of calculations, then undoing that normalisation to return to a coordinate-free perspective, doing some coordinate-free manipulations, and then performing a different normalisation to work on another part of the problem, and so forth. (For instance, in Euclidean geometry problems, it is often convenient to temporarily assign one key point to be the origin (thus spending translation invariance symmetry), then another, then switch back to a translation-invariant perspective, and so forth. As long as one is correctly accounting for what symmetries are being spent and bought at any given time, this can be a very powerful way of simplifying one's calculations.)

Given a coordinate system $\Phi: A \to G$ that identifies some geometric object A with a standard object G, and some isomorphism $\Psi: G \to G$ of that standard object, we can obtain a new coordinate system $\Psi \circ \Phi: A \to G$ of A by composing the two isomorphisms. [I will be vague on what "isomorphism" means; one can formalise the concept using the language of category theory.] Conversely, every other coordinate system $\Phi': A \to G$ of A arises in this manner. Thus, the space of coordinate systems on A is (non-canonically) identifiable with the isomorphism group $\mathrm{Isom}(G)$ of G. This isomorphism group is called the *structure group* (or *gauge group*) of the class of geometric objects. For example, the structure group for lengths is $\mathbb{R}^+$; the structure group for angles is $\mathbb{Z}/2\mathbb{Z}$; the structure group for lines is the affine group $\mathrm{Aff}(\mathbb{R})$; the structure group for n-dimensional Euclidean geometry is the Euclidean group $E(n)$; the structure group for (oriented) 2-spheres is the (special) orthogonal group $SO(3)$; and so forth. (Indeed, one can basically describe each of the classical geometries (Euclidean, affine, projective, spherical, hyperbolic, Minkowski, etc.) as a homogeneous space for its structure group, as per the Erlangen program.)

– Gauges –

In our discussion of coordinate systems, we focused on a single geometric (or combinatorial) object A: a single line, a single circle, a single set, etc. We then used a single coordinate system to identify that object with a standard representative of such an object.

Now let us consider the more general situation in which one has a *family* (or *fibre bundle*) $(A_x)_{x \in X}$ of geometric (or combinatorial) objects (or *fibres*) A_x: a family of lines (i.e. a line bundle), a family of circles (i.e. a circle bundle), a family of sets, etc. This family is parameterised by some *parameter set* or *base point* x, which ranges in some *parameter space* or *base space* X. In many cases one also requires some topological or differentiable compatibility between the various fibres; for instance, continuous (or smooth) variations of the base point should lead to continuous (or smooth) variations in the fibre. For sake of discussion, however, let us gloss over these compatibility conditions.

In many cases, each individual fibre A_x in a bundle $(A_x)_{x \in X}$, being a geometric object of a certain class, can be identified with a standard object G in that class, by means of a separate coordinate system $\Phi_x: A_x \to G$ for each base point x. The entire collection $\Phi = (\Phi_x)_{x \in X}$ is then referred to as a (global) *gauge* or *trivialisation* for this bundle (provided that it is compatible with whatever topological or differentiable structures one has placed on the bundle, but never mind that for now). Equivalently, a gauge is a bundle isomorphism Φ from the original bundle $(A_x)_{x\in X}$ to the *trivial bundle* $(G)_{x\in X}$, in which every fibre is the standard geometric object G. (There are also *local* gauges, which only trivialise a portion of the bundle, but let's ignore this distinction for now.)

Let's give three concrete examples of bundles and gauges; one from differential geometry, one from dynamical systems, and one from combinatorics.

Example 1: the circle bundle of the sphere. Recall from the previous section that the space of directions in a plane (which can be viewed as the circle of unit vectors) can be identified with the standard circle S^1 after picking an orientation and a reference direction. Now let us work not on the plane, but on a sphere, and specifically, on the surface X of the earth. At each point x on this surface, there is a circle S_x of directions that one can travel along the sphere from x; the collection $SX := (S_x)_{x\in X}$ of all such circles is then a circle bundle with base space X (known as *the* circle bundle; it could also be viewed as the sphere bundle, cosphere bundle, or orthonormal frame bundle of X). The structure group of this bundle is the circle group $U(1) \equiv S^1$ if one preserves orientation, or the semi-direct product $S^1 \rtimes \mathbb{Z}/2\mathbb{Z}$ otherwise.

Now suppose, at every point x on the earth X, the wind is blowing in some direction $w_x \in S_x$. (This is not actually possible globally, thanks to the hairy ball theorem, but let's ignore this technicality for now.) Thus wind direction can be thought of as a collection $w = (w_x)_{x\in X}$ of representatives from the fibres of the fibre bundle $(S_x)_{x\in X}$; such a collection is known as a *section* of the fibre bundle (it is to bundles as the concept of a graph $\{(x, f(x)) : x \in X\} \subset X \times G$ of a function $f : X \to G$ is to the trivial bundle $(G)_{x\in X}$).

At present, this section has not been represented in terms of numbers; instead, the wind direction $w(w_x)_{x\in X}$ is a collection of points on various different circles in the circle bundle SX. But one can convert this section w into a collection of numbers (and more specifically, a function $u : X \to S^1$ from X to S^1) by choosing a gauge for this circle bundle – in other words, by selecting an orientation ϵ_x and a reference direction N_x for each point x on the surface of the Earth X. For instance, one can pick the anticlockwise orientation $\circlearrowleft$ and true north for every point x (ignore for now the problem that this is not defined at the north and south poles, and so is merely a local gauge rather than a global one), and then each wind direction w_x can now be identified with a unit complex number $u(x) \in S^1$ (e.g. $e^{i\pi/4}$ if the wind is blowing in the northwest direction at x). Now that one has a numerical function u to play with, rather than a geometric object w, one can now use analytical tools (e.g. differentiation, integration, Fourier transforms, etc.) to analyse the wind direction if one desires. But one should be aware that this function reflects the choice of gauge as well as the original object of study. If one changes the gauge (e.g. by using magnetic north instead of true north), then the function u changes, even though the wind direction w is still the same. If one does not want to spend the U(1) gauge symmetry, one would have to take care that all operations one performs on these functions are gauge-invariant; unfortunately, this restrictive requirement eliminates wide swathes of analytic tools (in particular, integration and the Fourier transform) and so one is often forced to break the gauge symmetry in order to use analysis. The challenge is then to select the gauge that maximises the effectiveness of analytic methods. $\diamond$

Example 2: circle extensions of a dynamical system. Recall (see e.g. my lecture notes) that a dynamical system is a pair X = (X,T), where X is a space and $T : X \to X$ is an invertible map. (One can also place additional topological or measure-theoretic structures on this system, as is done in those notes, but we will ignore these structures for this discussion.) Given such a system, and given a *cocycle* $\rho : X \to S^1$ (which, in this context, is simply a function from X to the unit circle), we can define the *skew product* $X \times_\rho S^1$ of X and the unit circle S^1, twisted by the cocycle ρ, to be the Cartesian product $X \times S^1 := \{(x,u) : x \in X, u \in S^1\}$ with the shift $\tilde T : (x,u) \mapsto (Tx, \rho(x)u)$; this is easily seen to be another dynamical system. (If one wishes to have a topological or measure-theoretic dynamical system, then ρ will have to be continuous or measurable here, but let us ignore such issues for this discussion.) Observe that there is a free action $(S_v : (x,u) \mapsto (x, vu))_{v \in S^1}$ of the circle group S^1 on the skew product $X \times_\rho S^1$ that commutes with the shift $\tilde T$; the quotient space $(X \times_\rho S^1)/S^1$ of this action is isomorphic to X, thus leading to a *factor map* $\pi : X \times_\rho S^1 \to X$, which is of course just the projection map $\pi : (x,u) \mapsto x$. (An example is provided by the *skew shift system*, described in my lecture notes.)

Conversely, suppose that one had a dynamical system $\tilde X = (\tilde X, \tilde T)$ which had a free S^1 action $(S_v : \tilde X \to \tilde X)_{v \in S^1}$ commuting with the shift $\tilde T$. If we set $X := \tilde X / S^1$ to be the quotient space, we thus have a factor map $\pi : \tilde X \to X$, whose level sets $\pi^{-1}(\{x\})$ are all isomorphic to the circle S^1; we call $\tilde X$ a *circle extension* of the dynamical system X. We can thus view $\tilde X$ as a *circle bundle* $(\pi^{-1}(\{x\}))_{x \in X}$ with base space X, thus the level sets $\pi^{-1}(\{x\})$ are now the fibres of the bundle, and the structure group is S^1. If one picks a *gauge* for this bundle, by choosing a reference point $p_x \in \pi^{-1}(\{x\})$ in the fibre for each base point x (thus in this context a gauge is the same thing as a *section* $p = (p_x)_{x \in X}$; this is basically because this bundle is a principal bundle), then one can identify $\tilde X$ with a skew product $X \times_\rho S^1$ by identifying the point $S_v p_x \in \tilde X$ with the point $(x,v) \in X \times_\rho S^1$ for all $x \in X, v \in S^1$, and letting ρ be the cocycle defined by the formula

$$S_{\rho(x)} p_{Tx} = \tilde T p_x.$$

One can check that this is indeed an isomorphism of dynamical systems; if all the various objects here are continuous (resp. measurable), then one also has an isomorphism of topological dynamical systems (resp. measure-preserving systems). Thus we see that gauges allow us to write circle extensions as skew products. However, more than one gauge is available for any given circle extension; two gauges $(p_x)_{x \in X}$, $(p'_x)_{x \in X}$ will give rise to two skew products $X \times_\rho S^1$, $X \times_{\rho'} S^1$ which are isomorphic but not identical. Indeed, if we let $v : X \to S^1$ be a rotation map that sends p_x to p'_x, thus $p'_x = S_{v(x)} p_x$, then we see that the two cocycles ρ' and ρ are related by the formula

$$\rho'(x) = v(Tx)^{-1} \rho(x) v(x). \quad (1)$$

Two cocycles that obey the above relation are called *cohomologous*; their skew products are isomorphic to each other. An important general question in dynamical systems is to understand when two given cocycles are in fact cohomologous, for instance by introducing non-trivial cohomological invariants for such cocycles.

As an example of a circle extension, consider the sphere $X = S^2$ from Example 1, with a rotation shift T given by, say, rotating anti-clockwise by some given angle α around the axis connecting the north and south poles. This rotation also induces a rotation on the circle bundle $\tilde X := SX$, thus giving a circle extension of the original system (X,T). One can then use a gauge to write this system as a skew product. For instance, if one selects the gauge that chooses p_x to be the true north direction at each point x (ignoring for now the fact that this is not defined at the two poles), then this system becomes the ordinary product $X \times_0 S^1$ of the original system X with the circle S^1, with the cocycle being the trivial cocycle 0. If we were however to use a different gauge, e.g. magnetic north instead of true north, one would obtain a different skew-product $X \times_{\rho'} S^1$, where ρ' is some cocycle which is cohomologous to the trivial cocycle (except at the poles). (A cocycle which is globally cohomologous to the trivial cocycle is known as a *coboundary*. Not every cocycle is a coboundary, especially once one imposes topological or measure-theoretic structure, thanks to the presence of various topological or measure-theoretic invariants, such as degree.)

There was nothing terribly special about circles in this example; one can also define group extensions, or more generally homogeneous space extensions, of dynamical systems, and have a similar theory, although one has to take a little care with the order of operations when the structure group is non-abelian; see e.g. my lecture notes on isometric extensions. $\diamond$

Example 3: Orienting an undirected graph. The language of gauge theory is not often used in combinatorics, but nevertheless combinatorics does provide some simple discrete examples of bundles and gauges which can be useful in getting an intuitive grasp of the concept. Consider for instance an undirected graph G = (V,E) of vertices and edges. I will let X=E denote the space of edges (not the space of vertices)!. Every edge $e \in X$ can be oriented (or directed) in two different ways; let A_e be the pair of directed edges of e arising in this manner. Then $(A_e)_{e\in X}$ is a fibre bundle with base space X and with each fibre isomorphic (in the category of sets) to the standard two-element set $\{-1,+1\}$, with structure group $\mathbb{Z}/2\mathbb{Z}$.

A priori, there is no reason to prefer one orientation of an edge e over another, and so there is no canonical way to identify each fibre A_e with the standard set $\{-1,+1\}$. Nevertheless, we can go ahead and arbitrary select a gauge for X by *orienting* the graph G. This orientation assigns an oriented edge $\vec{e} \in A_e$ to each edge $e \in X$, thus creating a gauge (or section) $(\vec{e})_{e\in X}$ of the bundle $(A_e)_{e\in X}$. Once one selects such a gauge, we can now identify the fibre bundle $(A_e)_{e\in X}$ with the trivial bundle $X \times \{-1,+1\}$ by identifying the preferred oriented edge $\vec{e}$ of each unoriented edge $e \in X$ with $(e,+1)$, and the other oriented edge with $(e,-1)$. In particular, any other orientation of the graph G can be expressed relative to this reference orientation as a function $f: X \to \{-1,+1\}$, which measures when the two orientations agree or disagree with each other. $\diamond$

Recall that every isomorphism $\Psi \in \mathrm{Isom}(G)$ of a standard geometric object G allowed one to transform a coordinate system $\Phi: A \to G$ on a geometric object A to another coordinate system $\Psi \circ \Phi: A \to G$. We can generalise this observation to gauges: every family $\Psi = (\Psi_x)_{x\in X}$ of isomorphisms on G allows one to transform a gauge $(\Phi_x)_{x\in X}$ to another gauge $(\Psi_x \circ \Phi_x)_{x\in X}$ (again assuming that Ψ respects whatever topological or differentiable structure is present). Such a collection Ψ is known as a *gauge transformation*. For instance, in Example 1, one could rotate the reference direction N_x at each point $x \in X$ anti-clockwise by some angle $\theta(x)$; this would cause the function $u(x)$ to rotate to $u(x)e^{-i\theta(x)}$. In Example 2, a gauge transformation is just a map $v: X \to S^1$ (which may need to be continuous or measurable, depending on the structures one places on X); it rotates a point $(x,u) \in X \times_\rho S^1$ to $(x, v^{-1}u)$, and it also transforms the cocycle ρ by the formula (1). In Example 3, a gauge transformation would be a map $v: X \to \{-1,+1\}$; it rotates a point $(x,\epsilon) \in X \times \{-1,+1\}$ to $(x, v(x)\epsilon)$.

Gauge transformations transform functions on the base X in many ways, but some things remain gauge-invariant. For instance, in Example 1, the winding number of a function $u: X \to S^1$ along a closed loop

$\gamma \subset X$ would not change under a gauge transformation (as long as no singularities in the gauge are created, moved, or destroyed, and the orientation is not reversed). But such topological gauge-invariants are not the only gauge invariants of interest; there are important *differential* gauge-invariants which make gauge theory a crucial component of modern differential geometry and geometric PDE. But to describe these, one needs an additional gauge-theoretic concept, namely that of a *connection* on a fibre bundle.

– Connections –

There are many essentially equivalent ways to introduce the concept of a connection; I will use the formulation based primarily on parallel transport, and on differentiation of sections. To avoid some technical details I will work (somewhat non-rigorously) with infinitesimals such as dx. (There are ways to make the use of infinitesimals rigorous, such as non-standard analysis, but this is not the focus of my post today.)

In single variable calculus, we learn that if we want to differentiate a function $f: [a,b] \to \mathbb{R}$ at some point x, then we need to compare the value f(x) of f at x with its value f(x+dx) at some infinitesimally close point x+dx, take the difference $f(x+dx) - f(x)$, and then divide by dx, taking limits as $dx \to 0$, if one does not like to use infinitesimals:

$$\nabla f(x) := \lim_{dx \to 0} \frac{f(x+dx) - f(x)}{dx}.$$

In several variable calculus, we learn several generalisations of this concept in which the domain and range of f to be multi-dimensional. For instance, if $f: X \to \mathbb{R}^d$ is now a vector-valued function on some multi-dimensional domain (e.g. a manifold) X, and v is a tangent vector to X at some point x, we can define the directional derivative $\nabla_v f(x)$ of f at x by comparing $f(x+vdt)$ with $f(x)$ for some infinitesimal dt, take the difference $f(x+vdt) - f(x)$, divide by dt, and then take limits as $dt \to 0$:

$$\nabla_v f(x) := \lim_{dt \to 0} \frac{f(x+vdt) - f(x)}{dt}.$$

[Strictly speaking, if X is not flat, then x+vdt is only defined up to an ambiguity of o(dt), but let us ignore this minor issue here, as it is not important in the limit.] If f is sufficiently smooth (being continuously differentiable will do), the directional derivative is linear in v, thus for instance $\nabla_{v+v'} f(x) = \nabla_v f(x) + \nabla_{v'} f(x)$. One can also generalise the range of f to other multi-dimensional domains than $\mathbb{R}^d$; the directional derivative then lives in a tangent space of that domain.

In all of the above examples, though, we were differentiating functions $f: X \to Y$, thus each element $x \in X$ in the base (or domain) gets mapped to an element $f(x)$ in the same range Y. However, in many geometrical situations we would like to differentiate *sections* $f = (f_x)_{x \in X}$ instead of functions, thus f now maps each point $x \in X$ in the base to an element $f_x \in A_x$ of some fibre in a fibre bundle $(A_x)_{x \in X}$. For instance, one might want to know how the wind direction $w = (w_x)_{x \in X}$ changes as one moves x in some direction v; thus computing a directional derivative $\nabla_v w(x)$ of w at x in direction v. One can try to mimic the previous definitions in order to define this directional derivative. For instance, one can move x along v by some infinitesimal amount dt, creating a nearby point $x + vdt$, and then evaluate w at this point to obtain $w(x+vdt)$. But here we hit a snag: we cannot directly compare $w(x+vdt)$ with $w(x)$, because the former lives in the fibre A_{x+vdt} while the latter lives in the fibre A_x.

With a gauge, of course, we can identify all the fibres (and in particular, A_{x+vdt} and A_x) with a common object G, in which case there is no difficulty comparing $w(x+vdt)$ with $w(x)$. But this would lead to a notion of derivative which is not gauge-invariant, known as the *non-covariant* or *ordinary* derivative in physics.

But there is another way to take a derivative, which does not require the full strength of a gauge (which identifies *all* fibres simultaneously together). Indeed, in order to compute a derivative $\nabla_v w(x)$, one only needs to identify (or *connect*) two infinitesimally close fibres together: A_x and A_{x+vdt}. In practice, these two fibres are already "within O(dt) of each other" in some sense, but suppose in fact that we have some means $\Gamma(x \to x+vdt) : A_x \to A_{x+vdt}$ of identifying these two fibres together. Then, we can pull back $w(x+vdt)$ from A_{x+vdt} to A_x through $\Gamma(x \to x+vdt)$ to define the covariant derivative:

$$\nabla_v w(x) := \lim_{dt \to 0} \frac{\Gamma(x \to x+vdt)^{-1}(w(x+vdt)) - w(x)}{dt}.$$

In order to retain the basic property that $\nabla_v w$ is linear in v, and to allow one to extend the infinitesimal identifications $\Gamma(x \to x+dx)$ to non-infinitesimal identifications, we impose the property that the $\Gamma(x \to x+dx)$ to be approximately transitive in that

$$\Gamma(x+dx \to x+dx+dx') \circ \Gamma(x \to x+dx) \approx \Gamma(x \to x+dx+dx') \quad (1)$$

for all x, dx, dx', where the $\approx$ symbol indicates that the error between the two sides is o(|dx| + |dx'|). [The precise nature of this error is actually rather important, being essentially the *curvature* of the connection Γ at x in the directions dx, dx', but let us ignore this for now.] To oversimplify a little bit, any collection Γ of infinitesimal maps $\Gamma(x \to x+dx)$ obeying this property (and some technical regularity properties) is a *connection*.

[There are many other important ways to view connections, for instance the Christoffel symbol perspective that we will discuss a bit later. Another approach is to focus on the differentiation operation ∇_v rather than the identifications $\Gamma(x \to x+dx)$ or $\Gamma(\gamma)$, and in particular on the algebraic properties of this operation, such as linearity in v or derivation-type properties (in particular, obeying various variants of the Leibnitz rule). This approach is particularly important in algebraic geometry, in which the notion of an infinitesimal or of a path may not always be obviously available, but we will not discuss it here.]

The way we have defined it, a connection is a means of identifying two infinitesimally close fibres A_x, A_{x+dx} of a fibre bundle $(A_x)_{x \in X}$. But, thanks to (1), we can also identify two distant fibres A_x, A_y, provided that we have a path $\gamma : [a,b] \to X$ from $x = \gamma(a)$ to $y = \gamma(b)$, by concatenating the infinitesimal identifications by a non-commutative variant of a Riemann sum:

$$\Gamma(\gamma) := \lim_{\sup|t_{i+1}-t_i| \to 0} \Gamma(\gamma(t_{n-1}) \to \gamma(t_n)) \circ \ldots \circ \Gamma(\gamma(t_0) \to \gamma(t_1)), \quad (2)$$

where $a = t_0 < t_1 < \ldots < t_n = b$ ranges over partitions. This gives us a *parallel transport* map $\Gamma(\gamma) : A_x \to A_y$ identifying A_x with A_y, which in view of its Riemann sum definition, can be viewed as the "integral" of the connection Γ along the curve γ. This map does not depend on how one parametrises the path γ, but it can depend on the choice of path used to travel from x to y.

We illustrate these concepts using several examples, including the three examples introduced earlier.

Example 1 continued. (Circle bundle of the sphere) The geometry of the sphere X in Example 1 provides a natural connection on the circle bundle SX, the Levi-Civita connection Γ, that lets one transport directions around the sphere in as "parallel" a manner as possible; the precise definition is a little technical (see e.g. my lecture notes for a brief description). Suppose for instance one starts at some location x on the equator of the earth, and moves to the antipodal point y by a great semi-circle γ going through the north pole. The parallel transport $\Gamma(\gamma): S_x \to S_y$ along this path will map the north direction at x to the *south* direction at y. On the other hand, if we went from x to y by a great semi-circle γ' going along the equator, then the north direction at x would be transported to the *north* direction at y. Given a section u of this circle bundle, the quantity $\nabla_v u(x)$ can be interpreted as the rate at which u rotates as one travels from x with velocity v. $\diamond$

Example 2 continued. (Circle extensions) In Example 2, we change the notion of "infinitesimally close" by declaring x and Tx to be infinitesimally close for any x in the base space X (and more generally, x and $T^n x$ are non-infinitesimally close for any positive integer n, being connected by the path $x \to Tx \to \ldots \to T^n x$, and similarly for negative n). A cocycle $\rho: X \to S^1$ can then be viewed as defining a connection on the skew product $X \times_\rho S^1$, by setting $\Gamma(x \mapsto Tx) = \rho(x)$ (and also $\Gamma(x \to x) = 1$ and $\Gamma(Tx \to x) = \rho(x)^{-1}$ to ensure compatibility with (1); to avoid notational ambiguities let us assume for sake of discussion that $x, Tx, T^{-1}x$ are always distinct from each other). The non-infinitesimal connections $\rho_n(x) := \Gamma(x \to Tx \to \ldots \to T^n x)$ are then given by the formula $\rho_n(x) = \rho(x)\rho(Tx)\ldots\rho(T^{n-1}x)$ for positive n (with a similar formula for negative n). Note that these iterated cocycles ρ_n also describe the iterations of the shift $\tilde T: (x,u) \mapsto (Tx, \rho(x)u)$, indeed $\tilde T^n(x,u) = (T^n x, \rho_n(x)u)$. $\diamond$

Example 3 continued. (Oriented graphs) In Example 3, we declare two edges e, e' in X to be "infinitesimally close" if they are adjacent. Then there is a natural notion of parallel transport on the bundle $(A_e)_{e \in X}$; given two adjacent edges $e = \{u,v\}, e' = \{v,w\}$, we let $\Gamma(e \to e')$ be the isomorphism from $A_e = \{\vec{uv}, \vec{vu}\}$ to $A_{e'} = \{\vec{vw}, \vec{wv}\}$ that maps $\vec{uv}$ to $\vec{vw}$ and $\vec{vu}$ to $\vec{wv}$. Any path $\gamma = (\{v_1,v_2\}, \{v_2,v_3\}, \ldots, \{v_{n-1},v_n\})$ of edges then gives rise to a connection $\Gamma(\gamma)$ identifying $A_{\{v_1,v_2\}}$ with $A_{\{v_{n-1},v_n\}}$. For instance, the triangular path $(\{u,v\},\{v,w\},\{w,u\},\{u,v\})$ induces the identity map on $A_{\{u,v\}}$, whereas the U-turn path $(\{u,v\},\{v,w\},\{w,x\},\{x,v\},\{v,u\})$ induces the anti-identity map on $A_{\{u,v\}}$.

Given an orientation $\vec G = (\vec e)_{e \in X}$ of the graph G, one can "differentiate" $\vec G$ at an edge $\{u,v\}$ in the direction $\{u,v\} \to \{v,w\}$ to obtain a number $\nabla_{\{u,v\}\to\{v,w\}} \vec G(\{u,v\}) \in \{-1,+1\}$, defined as +1 if the parallel transport from $\{u,v\}$ and $\{v,w\}$ preserves the orientations given by $\vec G$, and -1 otherwise. This number of course depends on the choice of orientation. But certain combinations of these numbers are independent of such a choice; for instance, given any closed path $\gamma = \{e_1, e_2, \ldots, e_n, e_{n+1} = e_1\}$ of edges in X, the "integral" $\prod_{i=1}^n \nabla_{e_i \to e_{i+1}} \vec G(e_i) \in \{-1,+1\}$ is independent of the choice of orientation $\vec G$ (indeed, it is equal to +1 if $\Gamma(\gamma)$ is the identity, and -1 if $\Gamma(\gamma)$ is the anti-identity. $\diamond$

Example 4. (Monodromy) One can interpret the monodromy maps of a covering space in the language of connections. Suppose for instance that we have a covering space $\pi: \tilde X \to X$ of a topological space X whose fibres $\pi^{-1}(\{x\})$ are discrete; thus $\tilde X$ is a discrete fibre bundle over X. The discreteness induces a natural connection Γ on this space, which is given by the lifting map; in particular, if one integrates this connection on a closed loop based at some point x, one obtains the monodromy map of that loop at x. $\diamond$

Example 5. (Definite integrals) In view of the definition (2), it should not be surprising that the definite integral $\int_a^b f(x)\ dx$ of a scalar function $f: [a,b] \to \mathbb{R}$ can be interpreted as an integral of a connection. Indeed, set $X := [a,b]$, and let $(\mathbb{R})_{x \in X}$ be the trivial line bundle over X. The function f induces a connection Γ_f on this bundle by setting

$$\Gamma_f(x \mapsto x+dx) : y \mapsto y + f(x) dx.$$

The integral $\Gamma_f([a,b])$ of this connection along $[a,b]$ is then just the operation of translation by $\int_a^b f(x)\ dx$ in the real line. $\diamond$

Example 6. (Line integrals) One can generalise Example 5 to encompass line integrals in several variable calculus. Indeed, if X is an n-dimensional domain, then a vector field $f = (f_1,\ldots,f_n): X \to \mathbb{R}^n$ induces a connection Γ_f on the trivial line bundle $(\mathbb{R})_{x \in X}$ by setting

$$\Gamma_f(x \mapsto x+dx) : y \mapsto y + f_1(x) dx_1 + \ldots + f_n(x) dx_n.$$

The integral $\Gamma_f(\gamma)$ of this connection along a curve γ is then just the operation of translation by the line integral $\int_\gamma f \cdot dx$ in the real line.

Note that a gauge transformation in this context is just a vertical translation $(x,y) \mapsto (x, y+V(x))$ of the bundle $(\mathbb{R})_{x \in X} \equiv X \times \mathbb{R}$ by some potential function $V: X \to \mathbb{R}$, which we will assume to be smooth for sake of discussion. This transformation conjugates the connection Γ_f to the connection $\Gamma_{f - \nabla V}$. Note that this is a conservative transformation: the integral of a connection along a closed loop is unchanged by gauge transformation. $\diamond$

Example 7. (ODE) A different way to generalise Example 5 can be obtained by using the fundamental theorem of calculus to interpret $\int_{[a,b]} f(x)\ dx$ as the final value $u(b)$ of the solution to the initial value problem

$$u'(t) = f(t); \quad u(a) = 0$$

for the ordinary differential equation $u' = f$. More generally, the solution u(b) to the initial value problem

$$u'(t) = F(t, u(t)); \quad u(a) = u_0$$

for some $u: [a,b] \to \mathbb{R}^n$ taking values in some manifold Y, where $F: [a,b] \times \mathbb{R}^n \to \mathbb{R}^n$ is a function (let us take it to be Lipschitz, to avoid technical issues), can also be interpreted as the integral of a connection Γ on the trivial vector space bundle $(\mathbb{R}^n)_{t \in [a,b]}$, defined by the formula

$$\Gamma(t \mapsto t+dt) : y \mapsto y + F(t,y) dt.$$

Then $\Gamma[a,b]$ will map u_0 to $u(b)$, this is nothing more than the Euler method for solving ODE. Note that the method of integrating factors in solving ODE can be interpreted as an attempt to simplify the connection Γ via a gauge transformation. Indeed, it can be profitable to view the entire theory of connections as a multidimensional "variable-coefficient" generalisation of the theory of ODE. $\diamond$

Once one selects a gauge, one can express a connection in terms of that gauge. In the case of vector bundles (in which every fibre is a d-dimensional vector space for some fixed d), the covariant derivative $\nabla_v w(x)$ of a section w of that bundle along some vector v emanating from x can be expressed in any given gauge by the formula

$$\nabla_v w(x)^i = v^\alpha \partial_\alpha w(x)^i + v^\alpha \Gamma^i_{\alpha j} w(x)^j$$

where we use the gauge to express w(x) as a vector $(w(x)^1, \ldots, w(x)^d)$, the indices $i, j = 1, \ldots, d$ are summed over the fibre dimensions (and α summed over the base dimensions) as per the usual conventions, and the $\Gamma^i_{\alpha j} := (\nabla_{e_\alpha} e_j)^i$ are the Christoffel symbols of this connection relative to this gauge.

One example of this, which models electromagnetism, is a connection on a complex line bundle $V = (V_{t,x})_{(t,x) \in \mathbb{R}^{1+3}}$ in spacetime $\mathbb{R}^{1+3} = \{(t,x) : t \in \mathbb{R}, x \in \mathbb{R}^3\}$. Such a bundle assigns a complex line $V_{t,x}$ (i.e. a one-dimensional complex vector space, and thus isomorphic to $\mathbb{C}$) to every point (t,x) in spacetime. The structure group here is U(1) (strictly speaking, this means that we view the fibres as *normed* one-dimensional complex vector spaces, otherwise the structure group would be $\mathbb{C}^\times$). A gauge identifies V with the trivial complex line bundle $(\mathbb{C})_{(t,x) \in \mathbb{R}^{1+3}}$, thus converting sections $(w_{t,x})_{(t,x) \in \mathbb{R}^{1+3}}$ of this bundle into complex-valued functions $\phi : \mathbb{R}^{1+3} \to \mathbb{C}$. A connection on V, when described in this gauge, can be given in terms of fields $A_\alpha : \mathbb{R}^{1+3} \to \mathbb{R}$ for $\alpha = 0, 1, 2, 3$; the covariant derivative of a section in this gauge is then given by the formula

$$\nabla_\alpha \phi := \partial_\alpha \phi + i A_\alpha \phi.$$

In the theory of electromagnetism, A_0 and (A_1, A_2, A_3) are known (up to some normalising constants) as the electric potential and magnetic potential respectively. Sections of V do not show up directly in Maxwell's equations of electromagnetism, but appear in more complicated variants of these equations, such as the Maxwell-Klein-Gordon equation.

A gauge transformation of V is given by a map $U : \mathbb{R}^{1+3} \to S^1$; it transforms sections by the formula $\phi \mapsto U^{-1}\phi$, and connections by the formula $\nabla_\alpha \mapsto U^{-1} \nabla_\alpha U$, or equivalently

$$A_\alpha \mapsto A_\alpha + \tfrac{1}{i} U^{-1} \partial_\alpha U = A_\alpha + \partial_\alpha \tfrac{1}{i} \log U. \quad (2)$$

In particular, the electromagnetic potential A_α is not gauge invariant (which broadly corresponds to the concept of being *nonphysical* or *nonmeasurable* in physics), as gauge symmetry allows one to add an arbitrary gradient function to this potential. However, the curvature tensor

$$F_{\alpha\beta} := [\nabla_\alpha, \nabla_\beta] = \partial_\alpha A_\beta - \partial_\beta A_\alpha$$

of the connection is gauge-invariant, and physically measurable in electromagnetism; the components $F_{0i} = -F_{i0}$ for $i = 1, 2, 3$ of this field have a physical interpretation as the electric field, and the components $F_{ij} = -F_{ji}$ for $1 \leq i < j \leq 3$ have a physical interpretation as the magnetic field. (The curvature tensor F can be interpreted as describing the parallel transport of infinitesimal rectangles; it measures how far off the connection is from being *flat*, which means that it can be (locally) "straightened" via some choice of gauge to be the trivial connection. In nonabelian gauge theories, in which the structure group is more complicated than just the abelian group U(1), the curvature tensor is non-scalar, but remains gauge-invariant in a tensor sense (gauge transformations will transform the curvature as they would transform a tensor of the same rank).

Gauge theories can often be expressed succinctly in terms of a connection and its curvatures. For instance, Maxwell's equations in free space, which describes how electromagnetic radiation propagates in the presence of charges and currents (but no media other than vacuum), can be written (after normalising away some physical constants) as

$$\partial^\alpha F_{\alpha\beta} = J_\beta$$

where J_β is the 4-current. (Actually, this is only half of Maxwell's equations, but the other half are a consequence of the interpretation (*) of the electromagnetic field as a curvature of a U(1) connection. Thus this purely geometric interpretation of electromagnetism has some non-trivial physical implications, for instance ruling out the possibility of (classical) magnetic monopoles.) If one generalises from complex line bundles to higher-dimensional vector bundles (with a larger structure group), one can then write down the (classical) Yang-Mills equation

$$\nabla^\alpha F_{\alpha\beta} = 0$$

which is the classical model for three of the four fundamental forces in physics: the electromagnetic, weak, and strong nuclear forces (with structure groups U(1), SU(2), and SU(3) respectively). (The classical model for the fourth force, gravitation, is given by a somewhat different geometric equation, namely the Einstein equations $G_{\alpha\beta} = 8\pi T_{\alpha\beta}$, though this equation is also "gauge-invariant" in some sense.)

The gauge invariance (or gauge freedom) inherent in these equations complicates their analysis. For instance, due to the gauge freedom (2), Maxwell's equations, when viewed in terms of the electromagnetic potential A_α, are ill-posed: specifying the initial value of this potential at time zero does not uniquely specify the future value of this potential (even if one also specifies any number of additional time derivatives of this potential at time zero), since one can use (2) with a gauge function U that is trivial at time zero but non-trivial at some future time to demonstrate the non-uniqueness. Thus, in order to use standard PDE methods to solve these equations, it is necessary to first fix the gauge to a sufficient extent that it eliminates this sort of ambiguity. If one were in a one-dimensional situation (as opposed to the four-dimensional situation of spacetime), with a trivial topology (i.e. the domain is a line rather than a circle), then it is possible to gauge transform the connection to be completely trivial, for reasons generalising both the fundamental theorem of calculus and the fundamental theorem of ODEs. (Indeed, to trivialise a connection Γ on a line $\mathbb{R}$, one can pick an arbitrary origin $t_0 \in \mathbb{R}$ and gauge transform each point $t \in \mathbb{R}$ by $\Gamma([t_0, t])$.) However, in higher dimensions, one cannot hope to completely trivialise a connection by gauge transforms (mainly because of the possibility of a non-zero curvature form); in general, one cannot hope to do much better than setting a single component of the connection to equal zero. For instance, for Maxwell's equations (or the Yang-Mills equations), one can trivialise the connection A_α in the time direction, leading to the *temporal gauge condition*

$$A_0 = 0.$$

This gauge is indeed useful for providing an easy proof of local existence for these equations, at least for smooth initial data. But there are many other useful gauges also that one can fix; for instance one has the Lorenz gauge

$$\partial^\alpha A_\alpha = 0$$

which has the nice property of being Lorentz-invariant, and transforms the Maxwell or Yang-Mills equations into linear or nonlinear wave equations respectively. Another important gauge is the Coulomb gauge

$$\partial_i A_i = 0$$

where i only ranges over spatial indices 1,2,3 rather than over spacetime indices 0,1,2,3. This gauge has an elliptic variational formulation (Coulomb gauges are critical points of the functional $\int_{\mathbf{R}^3} \sum_{i=1}^3 |A_i|^2$) and thus are expected to be "smaller" and "smoother" than many other gauges; this intuition can be borne out by standard elliptic theory (or Hodge theory, in the case of Maxwell's equations). In some cases, the correct selection of a gauge is crucial in order to establish basic properties of the underlying equation, such as local existence. For instance, the simplest proof of local existence of the Einstein equations uses a harmonic gauge, which is analogous to the Lorenz gauge mentioned earlier; the simplest proof of local existence of Ricci flow uses a gauge of de Turck that is also related to harmonic maps (see e.g. my lecture notes); and in my own work on wave maps, a certain "caloric gauge" based on harmonic map heat flow is crucial (see e.g. this post of mine). But in many situations, it is not yet fully understood whether the use of the correct choice of gauge is a mere technical convenience, or is more innate to the equation. It is definitely conceivable, for instance, that a given gauge field equation is well-posed with one choice of gauge but ill-posed with another. It would also be desirable to have a more gauge-invariant theory of PDEs that did not rely so heavily on gauge theory at all, but this seems to be rather difficult; many of our most powerful tools in PDE (for instance, the Fourier transform) are highly non-gauge-invariant, which makes it very inconvenient to try to analyse these equations in a purely gauge-invariant setting.

Share this: Print Email Share

Possibly related posts: (automatically generated)

- FOR POSTING: OO 617
- C. Castro: Generalized Gravity in Clifford Spaces, Vacuum Energy and Grand …
- Quantum Gravity via a Symmetric Gauge Theory?