What is a gauge?

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

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

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

— Coordinate systems —

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

Classical mathematics, such as practised by the ancient Greeks, could be lo
divided into two disciplines, *geometry* and *number theory*, where I use the
very broadly, to encompass all sorts of mathematics dealing with any sort o

The two disciplines are unified by the concept of a *coordinate system*, whicl to convert geometric objects to numeric ones or vice versa. The most well l example of a coordinate system is the Cartesian coordinate system for the p more generally for a Euclidean space), but this is just one example of many systems. For instance:

1. One can convert a length (of, say, an interval) into an (unsigned) real n vice versa, once one fixes a unit of length (e.g. the metre or the foot). the coordinate system is specified by the choice of length unit.
2. One can convert a displacement along a line into a (signed) real numb versa, once one fixes a unit of length *and* an orientation along that line case, the coordinate system is specified by the length unit together wi of orientation. Alternatively, one can replace the unit of length and the 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, o once one fixes a unit of length, an orientation along the line, *and* an or line. Equivalently, one can pick an origin O and a unit displacement ve coordinate system essentially identifies the original line with the stanc $\mathbb{R}$.
4. One can generalise these systems to higher dimensions. For instance, convert a displacement along a plane into a vector in $\mathbb{R}^2$, or vice versa, fixes two linearly independent displacement vectors e_1, e_2 (i.e. a basis) plane; the Cartesian coordinate system is just one special case of this scheme. Similarly, one can convert a position on a plane to a vector in picks a basis e_1, e_2 for that plane as well as an origin O, thus identifying 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) coor systems.)
5. To convert an angle in a plane to a signed number (modulo multiples c versa, one needs to pick an orientation on the plane (e.g. to decide tha clockwise angles are positive).
6. To convert a *direction* in a plane to a signed number (again modulo mu), or vice versa, one needs to pick an orientation on the plane, as well a reference direction (e.g. true or magnetic north is often used in the ca navigation).
7. Similarly, to convert a position on a circle to a number (modulo multipl vice versa, one needs to pick an orientation on that circle, together wi on that circle. Such a coordinate system then equates the original circ standard unit circle $S^1 := \{z \in \mathbb{C} : |z| = 1\}$ (with the standard origin $+1$ a standard anticlockwise orientation $\circlearrowleft$).
8. To convert a position on a two-dimensional sphere (e.g. the surface of

a first approximation) to a point on the standard unit sphere $S^2 := \{(x,y,z) \in \mathbb{R}^3 : x^2+y^2+z^2 = 1\}$, one can pick an orientation on tha "origin" (or "north pole") for that sphere, and a "prime meridian" conn north pole to its antipode. Alternatively, one can view this coordinate determining a pair of Euler angles ϕ, λ (or a latitude and longitude) to to every point on one's original sphere.

9. The above examples were all geometric in nature, but one can also co "combinatorial" coordinate systems, which allow one to identify combi objects with numerical ones. An extremely familiar example of this is *enumeration*: one can identify a set A of (say) five elements with the nu 1,2,3,4,5 simply by choosing an enumeration $a_1, a_2, \ldots, a_5$ of the set A. similarly enumerate other combinatorial objects (e.g. graphs, relations partial orders, etc.), and indeed this is done all the time in combinator Similarly for algebraic objects, such as cosets of a subgroup H (or mor torsors of a group G); one can identify such a coset with H itself by de element of that coset to be the "identity" or "origin".

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

Coordinate systems identify geometric or combinatorial objects with numer standard) ones, but in many cases, there is no natural (or canonical) choice identification; instead, one may be faced with a variety of coordinate systen equally valid. One can of course just fix one such system once and for all, i there is no real harm in thinking of the geometric and numeric objects as b equivalent. If however one plans to change from one system to the next (or using such systems altogether), then it becomes important to carefully disti two types of objects, to avoid confusion. For instance, if an interval AB is m have a length of 3 yards, then it is OK to write $|AB| = 3$ (identifying the geor concept of length with the numeric concept of a positive real number) so lo plan to stick to having the yard as the unit of length for the rest of one's an one was also planning to use, say, feet, as a unit of length also, then to avoi statements such as "$|AB| = 3$ and $|AB| = 9$", one should specify the coordina explicitly, e.g. "$|AB| = 3 \hbox{ yards}$ and $|AB| = 9 \hbox{ feet}$". Similarly, identifying a poin with its coordinates (e.g. $P = (4,3)$) is safe as long as one intends to only use coordinate system throughout; but if one intends to change coordinates at s (or to switch to a coordinate-free perspective) then one should be more car

writing $P = 4e_1 + 3e_2$, or even $P = O + 4e_1 + 3e_2$, if the origin O and basis vect
one's coordinate systems might be subject to future change.

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

The above abstract *coordinate-free perspective* is equivalent to a more con
coordinate-invariant perspective, in which we do allow the use of coordinat
all geometric quantities to numeric ones, but insist that every statement th
down is invariant under changes of coordinates. For instance, if we shrink
unit of length by a factor $\lambda > 0$, then the numerical length of every interval i
a factor of λ, e.g. $|AB| \mapsto \lambda|AB|$. The coordinate-invariant approach to lengt
measurement then treats lengths such as $|AB|$ as numbers, but requires all
involving such lengths to be invariant under the above scaling symmetry. F
a statement such as $|AC|^2 = |AB|^2 + |BC|^2$ is legitimate under this perspectiv
statement such as $|AB| = |BC|^2$ or $|AB| = 3$ is not. [In other words, co-ordina
invariance here is the same thing as being dimensionally consistent. Indee
[dimensional analysis](#) is nothing more than the analysis of the scaling symm
one's coordinate systems.] One can retain this coordinate-invariance symm
throughout one's arguments; or one can, at some point, choose to *spend* (o
coordinate invariance by selecting (or *fixing*) the coordinate system (which,
means selecting a unit length). The advantage in spending such a symmetr
can often normalise one or more quantities to equal a particularly nice valu
instance, if a length $|AB|$ is appearing everywhere in one's arguments, and
carefully retained coordinate-invariance up until some key point, then it ca

convenient to spend this invariance to normalise $|AB|$ to equal 1. (In this ca
has a one-dimensional family of symmetries, and so can only normalise one
time; but when one's symmetry group is larger, one can often normalise ma
quantities at once; as a rule of thumb, one can normalise one quantity for e
of freedom in the symmetry group.) Conversely, if one has already spent th
invariance, one can often buy it back by converting all the facts, hypotheses
desired conclusions one currently possesses in the situation back to a coord
invariant formulation. Thus one could imagine performing one normalisatic
set of calculations, then undoing that normalisation to return to a coordinat
perspective, doing some coordinate-free manipulations, and then performin
normalisation to work on another part of the problem, and so forth. (For in
Euclidean geometry problems, it is often convenient to temporarily assign c
to be the origin (thus spending translation invariance symmetry), then anot
switch back to a translation-invariant perspective, and so forth. As long as
correctly accounting for what symmetries are being spent and bought at an
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
standard object G, and some isomorphism $\Psi : G \to G$ of that standard object
obtain a new coordinate system $\Psi \circ \Phi : A \to G$ of A by composing the two isc
[I will be vague on what "isomorphism" means; one can formalise the conce
language of category theory.] Conversely, every other coordinate system Φ'
arises in this manner. Thus, the space of coordinate systems on A is (non-ca
identifiable with the isomorphism group $\mathrm{Isom}(G)$ of G. This isomorphism grc
the *structure group* (or *gauge group*) of the class of geometric objects. For
the structure group for lengths is $\mathbb{R}^+$; the structure group for angles is $\mathbb{Z}/2\mathbb{Z}$
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
(oriented) 2-spheres is the (special) orthogonal group $SO(3)$; and so forth. (
can basically describe each of the classical geometries (Euclidean, affine, p
spherical, hyperbolic, Minkowski, etc.) as a homogeneous space for its stru
as per the Erlangen program.)

— Gauges —

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

Now let us consider the more general situation in which one has a *family* (o
bundle) $(A_x)_{x \in X}$ of geometric (or combinatorial) objects (or *fibres*) A_x: a fam
(i.e. a line bundle), a family of circles (i.e. a circle bundle), a family of sets, (

family is parameterised by some *parameter set* or *base point* x, which range *parameter space* or *base space* X. In many cases one also requires some to differentiable compatibility between the various fibres; for instance, contin smooth) variations of the base point should lead to continuous (or smooth) the fibre. For sake of discussion, however, let us gloss over these compatib conditions.

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

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

Example 1: the circle bundle of the sphere. Recall from the previous se the space of directions in a plane (which can be viewed as the circle of unit be identified with the standard circle S^1 after picking an orientation and a r direction. Now let us work not on the plane, but on a sphere, and specifica surface X of the earth. At each point x on this surface, there is a circle S_x o that one can travel along the sphere from x; the collection $SX := (S_x)_{x\in X}$ of circles is then a circle bundle with base space X (known as *the* circle bundl also be viewed as the sphere bundle, cosphere bundle, or orthonormal fram X). The structure group of this bundle is the circle group $U(1) \equiv S^1$ if one pr 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 $w_x \in S_x$. (This is not actually possible globally, thanks to the hairy ball theor ignore this technicality for now.) Thus wind direction can be thought of as $w = (w_x)_{x\in X}$ of representatives from the fibres of the fibre bundle $(S_x)_{x\in X}$; su collection is known as a *section* of the fibre bundle (it is to bundles as the c graph $\{(x, f(x)) : x \in X\} \subset X \times G$ of a function $f : X \to G$ is to the trivial bund

At present, this section has not been represented in terms of numbers; inste wind direction $w(w_x)_{x\in X}$ is a collection of points on various different circles i bundle SX. But one can convert this section w into a collection of numbers specifically, a function $u : X \to S^1$ from X to S^1) by choosing a gauge for this – in other words, by selecting an orientation ϵ_x and a reference direction N_x

point x on the surface of the Earth X. For instance, one can pick the anticlc orientation $\circlearrowleft$ and true north for every point x (ignore for now the problem t not defined at the north and south poles, and so is merely a local gauge ratl global one), and then each wind direction w_x can now be identified with a u number $u(x) \in S^1$ (e.g. $e^{i\pi/4}$ if the wind is blowing in the northwest direction that one has a numerical function u to play with, rather than a geometric ol can now use analytical tools (e.g. differentiation, integration, Fourier transf to analyse the wind direction if one desires. But one should be aware that t reflects the choice of gauge as well as the original object of study. If one ch gauge (e.g. by using magnetic north instead of true north), then the functio even though the wind direction w is still the same. If one does not want to U(1) gauge symmetry, one would have to take care that all operations one p these functions are gauge-invariant; unfortunately, this restrictive requirem eliminates wide swathes of analytic tools (in particular, integration and the transform) and so one is often forced to break the gauge symmetry in order analysis. The challenge is then to select the gauge that maximises the effec analytic methods. $\diamond$

Example 2: circle extensions of a dynamical system. Recall (see e.g. m notes) that a dynamical system is a pair X = (X,T), where X is a space and T an invertible map. (One can also place additional topological or measure-th structures on this system, as is done in those notes, but we will ignore these for this discussion.) Given such a system, and given a *cocycle* $\rho : X \to S^1$ (w context, is simply a function from X to the unit circle), we can define the *sk* $X \times_\rho S^1$ of X and the unit circle S^1, twisted by the cocycle ρ, to be the Cartes $X \times S^1 := \{(x,u) : x \in X, u \in S^1\}$ with the shift $\tilde T : (x,u) \mapsto (Tx, \rho(x)u)$; this is ea be another dynamical system. (If one wishes to have a topological or measu dynamical system, then ρ will have to be continuous or measurable here, bu 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 $X \times_\rho S^1$ th with the shift $\tilde T$; the quotient space $(X \times_\rho S^1)/S^1$ of this action is isomorphic leading to a *factor map* $\pi : X \times_\rho S^1 \to X$, which is of course just the projectic $\pi : (x,u) \mapsto x$. (An example is provided by the *skew shift system*, described i notes.)

Conversely, suppose that one had a dynamical system $\tilde X = (\tilde X, \tilde T)$ which had 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 b quotient space, we thus have a factor map $\pi : \tilde X \to X$, whose level sets $\pi^{-1}(\{$ isomorphic to the circle S^1; we call $\tilde X$ a *circle extension* of the dynamical sy can thus view $\tilde X$ as a *circle bundle* $(\pi^{-1}(\{x\}))_{x \in X}$ with base space X, thus the $\pi^{-1}(\{x\})$ are now the fibres of the bundle, and the structure group is S^1. If o *gauge* for this bundle, by choosing a reference point $p_x \in \pi^{-1}(\{x\})$ in the fibr base point x (thus in this context a gauge is the same thing as a *section* $p =$ is basically because this bundle is a principal bundle), then one can identify skew product $X \times_\rho S^1$ by identifying the point $S_v p_x \in \tilde X$ with the point $(x, v) \in$ 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 a various objects here are continuous (resp. measurable), then one also has a isomorphism of topological dynamical systems (resp. measure-preserving sy Thus we see that gauges allow us to write circle extensions as skew produc However, more than one gauge is available for any given circle extension; t $(p_x)_{x \in X}$, $(p'_x)_{x \in X}$ will give rise to two skew products $X \times_\rho S^1$, $X \times_{\rho'} S^1$ which ar but not identical. Indeed, if we let $v : X \to S^1$ be a rotation map that sends p $p'_x = S_{v(x)} p_x$, then we see that the two cocycles ρ' and ρ are related by the for

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

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

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

There was nothing terribly special about circles in this example; one can als group extensions, or more generally homogeneous space extensions, of dyn systems, and have a similar theory, although one has to take a little care wi of operations when the structure group is non-abelian; see e.g. my lecture n isometric extensions. ◇

Example 3: Orienting an undirected graph. The language of gauge theo often used in combinatorics, but nevertheless combinatorics does provide s discrete examples of bundles and gauges which can be useful in getting an grasp of the concept. Consider for instance an undirected graph G = (V,E) and edges. I will let X=E denote the space of edges (not the space of vertic edge $e \in X$ can be oriented (or directed) in two different ways; let A_e be the directed edges of e arising in this manner. Then $(A_e)_{e \in X}$ is a fibre bundle wi space X and with each fibre isomorphic (in the category of sets) to the stand 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 anotl there is no canonical way to identify each fibre A_e with the standard set $\{-1$ Nevertheless, we can go ahead and arbitrary select a gauge for X by *orienti* graph G. This orientation assigns an oriented edge $\vec{e} \in A_e$ to each edge $e \in$ creating a gauge (or section) $(\vec{e})_{e \in X}$ of the bundle $(A_e)_{e \in X}$. Once one selects gauge, we can now identify the fibre bundle $(A_e)_{e \in X}$ with the trivial bundle by identifying the preferred oriented edge $\vec{e}$ of each unoriented edge $e \in X$ and the other oriented edge with $(e, -1)$. In particular, any other orientation graph G can be expressed relative to this reference orientation as a functio $f : X \to \{-1, +1\}$, which measures when the two orientations agree or disag each other. ◇

Recall that every isomorphism $\Psi \in \mathrm{Isom}(G)$ of a standard geometric object G to transform a coordinate system $\Phi : A \to G$ on a geometric object A to anotl coordinate system $\Psi \circ \Phi : A \to G$. We can generalise this observation to gau family $\Psi = (\Psi_x)_{x \in X}$ of isomorphisms on G allows one to transform a gauge $($ another gauge $(\Psi_x \circ \Phi_x)_{x \in X}$ (again assuming that Ψ respects whatever topolo differentiable structure is present). Such a collection Ψ is known as a *gaug transformation*. For instance, in Example 1, one could rotate the reference at each point $x \in X$ anti-clockwise by some angle $\theta(x)$; this would cause the to rotate to $u(x)e^{-i\theta(x)}$. In Example 2, a gauge transformation is just a map (which may need to be continuous or measurable, depending on the structu places on X); it rotates a point $(x, u) \in X \times_\rho S^1$ to $(x, v^{-1}u)$, and it also transfor cocycle ρ by the formula (1). In Example 3, a gauge transformation would $v : X \to \{-1, +1\}$; 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, bu things remain gauge-invariant. For instance, in Example 1, the winding nu function $u : X \to S^1$ along a closed loop $\gamma \subset X$ would not change under a gau transformation (as long as no singularities in the gauge are created, moved destroyed, and the orientation is not reversed). But such topological gauge are not the only gauge invariants of interest; there are important *differenti* invariants which make gauge theory a crucial component of modern differe geometry and geometric PDE. But to describe these, one needs an addition theoretic concept, namely that of a [connection] on a fibre bundle.

— Connections —

There are many essentially equivalent ways to introduce the concept of a co will use the formulation based primarily on parallel transport, and on differ sections. To avoid some technical details I will work (somewhat non-rigorou infinitesimals such as dx. (There are ways to make the use of infinitesimals 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 functi $f : [a, b] \to \mathbb{R}$ at some point x, then we need to compare the value f(x) of f at x 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 no 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 concep the domain and range of f to be multi-dimensional. For instance, if $f : X \to$ vector-valued function on some multi-dimensional domain (e.g. a manifold) tangent vector to X at some point x, we can define the directional derivative at x by comparing $f(x + vdt)$ with $f(x)$ for some infinitesimal dt, take the diff $f(x + vdt) - f(x)$, 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 ambig but let us ignore this minor issue here, as it is not important in the limit.] I sufficiently smooth (being continuously differentiable will do), the direction is linear in v, thus for instance $\nabla_{v+v'} f(x) = \nabla_v f(x) + \nabla_{v'} f(x)$. One can also ge range of f to other multi-dimensional domains than $\mathbb{R}^d$; the directional deriv lives in a tangent space of that domain.

In all of the above examples, though, we were differentiating functions $f : X$ each element $x \in X$ in the base (or domain) gets mapped to an element $f(x)$ range Y. However, in many geometrical situations we would like to differen *sections* $f = (f_x)_{x \in X}$ instead of functions, thus f now maps each point $x \in X$ i an element $f_x \in A_x$ of some fibre in a fibre bundle $(A_x)_{x \in X}$. For instance, one to know how the wind direction $w = (w_x)_{x \in X}$ changes as one moves x in som thus computing a directional derivative $\nabla_v w(x)$ of w at x in direction v. One mimic the previous definitions in order to define this directional derivative. instance, one can move x along v by some infinitesimal amount dt, creating point $x + vdt$, and then evaluate w at this point to obtain $w(x + vdt)$. But here snag: we cannot directly compare $w(x + vdt)$ with $w(x)$, because the former li 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+} with a common object G, in which case there is no difficulty comparing $w(x$ $w(x)$. But this would lead to a notion of derivative which is not gauge-invari as the *non-covariant* or *ordinary* derivative in physics.

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

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

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

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

for all x, dx, dx', where the $\approx$ symbol indicates that the error between the tv o(|dx| + |dx'|). [The precise nature of this error is actually rather importan essentially the *curvature* of the connection Γ at x in the directions dx, dx', bu ignore this for now.] To oversimplify a little bit, any collection Γ of infinitesi $\Gamma(x \to x + dx)$ obeying this property (and some technical regularity propertie *connection*.

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

The way we have defined it, a connection is a means of identifying two infin close fibres A_x, A_{x+dx} of a fibre bundle $(A_x)_{x \in X}$. But, thanks to (1), we can als two distant fibres A_x, A_y, provided that we have a path $\gamma : [a,b] \to X$ from $x =$ $y = \gamma(b)$, by concatenating the infinitesimal identifications by a non-commuta 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 *parall* map $\Gamma(\gamma) : A_x \to A_y$ identifying A_x with A_y, which in view of its Riemann sum can be viewed as the "integral" of the connection Γ along the curve γ. This not depend on how one parametrises the path γ, but it can depend on the c used to travel from x to y.

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

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

Example 2 continued. (Circle extensions) In Example 2, we change the nc "infinitesimally close" by declaring x and Tx to be infinitesimally close for a base space X (and more generally, x and $T^n x$ are non-infinitesimally close fo positive integer n, being connected by the path $x \to Tx \to \ldots \to T^n x$, and sin negative n). A cocycle $\rho : X \to S^1$ can then be viewed as defining a connecti 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 ambi assume for sake of discussion that $x, Tx, T^{-1}x$ are always distinct from each non-infinitesimal connections $\rho_n(x) := \Gamma(x \to Tx \to \ldots \to T^n x)$ are then given formula $\rho_n(x) = \rho(x)\rho(Tx) \ldots \rho(T^{n-1}x)$ for positive n (with a similar formula fo n). Note that these iterated cocycles ρ_n also describe the iterations of the s $\tilde{T} : (x, u) \mapsto (Tx, \rho(x)u)$, 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 ed X to be "infinitesimally close" if they are adjacent. Then there is a natural r parallel transport on the bundle $(A_e)_{e \in X}$; given two adjacent edges $e = \{u, v\}$, we let $\Gamma(e \to e')$ be the isomorphism from $A_e = \{\vec{uv}, \vec{vu}\}$ to $A_{e'} = \{\vec{vw}, \vec{wv}\}$ that $\vec{vw}$ 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 gi connection $\Gamma(\gamma)$ identifying $A_{\{v_1, v_2\}}$ with $A_{\{v_{n-1}, v_n\}}$. For instance, the triangula $(\{u, v\}, \{v, w\}, \{w, u\}, \{u, v\})$ induces the identity map on $A_{\{u,v\}}$, whereas the U- $(\{u, v\}, \{v, w\}, \{w, x\}, \{x, v\}, \{v, u\})$ 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 $\{u, v\}$ in the direction $\{u, v\} \to \{v, w\}$ to obtain a number $\nabla_{\{u,v\} \to \{v,w\}} \vec{G}(\{u, v\}) \in$ defined as +1 if the parallel transport from $\{u, v\}$ and $\{v, w\}$ preserves the or given by $\vec{G}$, and -1 otherwise. This number of course depends on the choice orientation. But certain combinations of these numbers are independent of choice; for instance, given any closed path $\gamma = \{e_1, e_2, \ldots, e_n, e_{n+1} = e_1\}$ of edg "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 orient (indeed, it is equal to +1 if $\Gamma(\gamma)$ is the identity, and -1 if $\Gamma(\gamma)$ is the anti-ident

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

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

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

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

Example 6. (Line integrals) One can generalise Example 5 to encompass li in several variable calculus. Indeed, if X is an n-dimensional domain, then $f = (f_1, \ldots, f_n) : X \to \mathbb{R}^n$ induces a connection Γ_f on the trivial line bundle $(\mathbb{R})$ 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 operati 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 translatio $(x, y) \mapsto (x, y + V(x))$ of the bundle $(\mathbb{R})_{x \in X} \equiv X \times \mathbb{R}$ by some potential function which we will assume to be smooth for sake of discussion. This transforma conjugates the connection Γ_f to the connection $\Gamma_{f-\nabla V}$. Note that this is a co transformation: the integral of a connection along a closed loop is unchang transformation. $\diamond$

Example 7. (ODE) A different way to generalise Example 5 can be obtaine the fundamental theorem of calculus to interpret $\int_{[a,b]} f(x)\ dx$ as the final val 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(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 $F : [a, b] \times \mathbb{R}^n$ function (let us take it to be Lipschitz, to avoid technical issues), can also b as the integral of a connection Γ on the trivial vector space bundle $(\mathbb{R}^n)_{t \in [a,b]}$, 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 fo ODE. Note that the method of integrating factors in solving ODE can be in an attempt to simplify the connection Γ via a gauge transformation. Indeed profitable to view the entire theory of connections as a multidimensional "v coefficient" generalisation of the theory of ODE. $\diamond$

Once one selects a gauge, one can express a connection in terms of that ga case of vector bundles (in which every fibre is a d-dimensional vector space fixed d), the covariant derivative $\nabla_v w(x)$ of a section w of that bundle along 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 in $i, j = 1, \ldots, d$ are summed over the fibre dimensions (and α summed over the dimensions) as per the usual conventions, and the $\Gamma^i_{\alpha j} := (\nabla_{e_\alpha} e_j)^i$ are the Chı symbols of this connection relative to this gauge.

One example of this, which models electromagnetism, is a connection on a [cut off] 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 bu[cut off] a complex line $V_{t,x}$ (i.e. a one-dimensional complex vector space, and thus is [cut off] $\mathbb{C}$) to every point (t,x) in spacetime. The structure group here is U(1) (strict[cut off] this means that we view the fibres as *normed* one-dimensional complex vec[cut off] otherwise the structure group would be $\mathbb{C}^\times$). A gauge identifies V with the t[cut off] 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 [cut off] complex-valued functions $\phi : \mathbb{R}^{1+3} \to \mathbb{C}$. A connection on V, when described i[cut off] gauge, can be given in terms of fields $A_\alpha : \mathbb{R}^{1+3} \to \mathbb{R}$ for $\alpha = 0,1,2,3$; the cova[cut off] 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[cut off] constants) as the electric potential and magnetic potential respectively. Se[cut off] do not show up directly in Maxwell's equations of electromagnetism, but ap[cut off] more complicated variants of these equations, such as the Maxwell-Klein-G[cut off] equation.

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

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

In particular, the electromagnetic potential A_α is not gauge invariant (which[cut off] corresponds to the concept of being *nonphysical* or *nonmeasurable* in physi[cut off] gauge symmetry allows one to add an arbitrary gradient function to this po[cut off] 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 electrom[cut off] the components $F_{0i} = -F_{i0}$ for $i = 1,2,3$ of this field have a physical interpret[cut off] 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 interp[cut off] describing the parallel transport of infinitesimal rectangles; it measures ho[cut off] connection is from being *flat*, which means that it can be (locally) "straighte[cut off] some choice of gauge to be the trivial connection. In nonabelian gauge the[cut off] which the structure group is more complicated than just the abelian group [cut off] curvature tensor is non-scalar, but remains gauge-invariant in a tensor sens[cut off] transformations will transform the curvature as they would transform a ten[cut off] same rank).

Gauge theories can often be expressed succinctly in terms of a connection a[cut off] curvatures. For instance, Maxwell's equations in free space, which describ electromagnetic radiation propagates in the presence of charges and curre media other than vacuum), can be written (after normalising away some ph 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 other half are a consequence of the interpretation (*) of the electromagneti curvature of a U(1) connection. Thus this purely geometric interpretation c electromagnetism has some non-trivial physical implications, for instance ru possibility of (classical) magnetic monopoles.) If one generalises from com bundles to higher-dimensional vector bundles (with a larger structure grou 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 phys electromagnetic, weak, and strong nuclear forces (with structure groups U(and SU(3) respectively). (The classical model for the fourth force, gravitati by a somewhat different geometric equation, namely the Einstein equations though this equation is also "gauge-invariant" in some sense.)

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

$$A_0 = 0.$$

This gauge is indeed useful for providing an easy proof of local existence fo equations, at least for smooth initial data. But there are many other useful 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 Yang-Mills equations into linear or nonlinear wave equations respectively. A 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 in 0,1,2,3. This gauge has an elliptic variational formulation (Coulomb gauges points of the functional $\int_{\mathbb{R}^3} \sum_{i=1}^3 |A_i|^2$) and thus are expected to be "smaller" "smoother" than many other gauges; this intuition can be borne out by stan theory (or Hodge theory, in the case of Maxwell's equations). In some case correct selection of a gauge is crucial in order to establish basic properties underlying equation, such as local existence. For instance, the simplest pro existence of the Einstein equations uses a harmonic gauge, which is analog Lorenz gauge mentioned earlier; the simplest proof of local existence of Ric a gauge of de Turck that is also related to harmonic maps (see e.g. my lectu and in my own work on wave maps, a certain "caloric gauge" based on harr heat flow is crucial (see e.g. this post of mine). But in many situations, it is understood whether the use of the correct choice of gauge is a mere techni convenience, or is more innate to the equation. It is definitely conceivable, that a given gauge field equation is well-posed with one choice of gauge bu with another. It would also be desirable to have a more gauge-invariant the that did not rely so heavily on gauge theory at all, but this seems to be rath many of our most powerful tools in PDE (for instance, the Fourier transform non-gauge-invariant, which makes it very inconvenient to try to analyse the in a purely gauge-invariant setting.

47 comments

Comments feed fc

28 September, 2008 at 4:17 pm
John Sidles

Oh boy! As first-poster (at the time of writing, any this is my chance to express appreciation of Teren wonderful series of lectures.

This particular (gauge theory) lecture touches upon a topic of great practic engineering, namely what the lecture calls "the challenge [of selecting] the maximizes the effectiveness of analytic methods."

Very often in engineering, one has a global algebraic invariance that one wi promote to a local geometry invariance … and then link to conservation law be too much to hope for a few remarks on this topic?

Within the context of quantum simulation science, specifically within the (c problem of dynamically simulating open quantum systems, there is a concre of this kind of mathematical challenge.

Namely, there is a well-known global algebraic symmetry associated with "a the operator-product representation". It is natural to ask—without necessa clear idea of what the answer might be—what mathematical tools are avail promoting this global algebraic symmetry on linear quantum state-spaces, t geometric symmetry on nonlinear quantum state-spaces?

For all us engineers know, this is may be a well-established area of mathem perhaps not … and in either case the necessary ideas are perhaps not all th recognize.

As pretty much everyone appreciates, weblogs like Terence's are a wonderf for helping people get started on these lines of inquiry.

So please accept my thanks, for a half-hour of pure enjoyment, which left b reader's mind) the germ of new perspectives and new lines of inquiry that (may grow.

Reply

29 September, 2008 at 6:54 am
Terry Tao and gauge theories « The Gauge Connection

[…] I have found a beatiful post by Fields medallist, about gauge theo for a worthwhile reading. This pos elucidating and so well written that I thought it was […]

Reply

30 September, 2008 at 4:44 am
Pedro Lauridsen Ribeiro

Dear Prof. Tao,

First of all, as a mathematical physicist, I must (a
others)
thank you for your clear and precise post on this beautiful topic.
I have a few comments regarding the last paragraph.

Indeed, gauge-invariant lines of attack for analytical aspects of PDE's endowed with gauge invariance or, more generally, with some sort of constraint with respect to the Cauchy problem (i.e. the system is under-determined but the initial data cannot be arbitrary – it satisfies some relations given by components of the PDE system in such a way that this relations are guaranteed to be satisfied for all times due to the remaining (evolution) part of the system, once they are satisfied by the initial data. Gauge fixing gives a solution to the constraint part of the PDE system) are usually based on configuration space techniques. A recent prime example is the paper by Klainerman and Rodnianski (J. Hyperbolic Differ. Equ. 4 (2007), 401-433) on the construction of a Kirchoff-Sobolev parametrix for the wave equation and its use for an alternative, _gauge-inva
of the global well-posedness of the
Yang-Mills system established earlier by Eardley and Moncrief. This construction also holds in curved spacetimes, and it's based on modifying the original Kirchoff-Sobolev construction by replacing the spatial distance by another one, adapted to the geometry of a null (i.e. characteristic) foliation of the spacetime. The latter device is also remniscent of the landmark (and also configuration space-based) proof by Christodoulou and Klainerman of the global nonlinear stability of Minkowski spacetime w.r.t. the Einstein equations, which also makes use of yet another configuration space method of obtaining estimates, namely that of commuting vector fields.

The question of well-posedness of PDE systems endowed with gauge invariance is related to another deep problem, namely: What is the definition of hyperbolicity of a PDE system endowed with gauge invariance (or, more generally, constraints)? A tentative one, which works for many important examples (Einstein, Yang-Mills, Maxwell, etc.), would be: "A PDE system with constraints such that there is a solution of the latter (i.e., a gauge fixing prescription) which renders the 'reduced' system hyperbolic in the usual sense", but this is too loose. Another line of attack would be to add extra, auxiliary functions (fields) which correspond either to derivatives

of the fields and/or to "Lagrange multipliers" and demand that the enlarged system is hyperbolic, but this enlargement also seems to depend on the particular structure of the PDE system at hand.
It seems to me that a proper, _gauge-invariant_ definition of hyperbolicity, which is obviously important from a physical viewpoint, is crucial even to devise gauge-invariant analytical tools in a more systematic fashion.

Reply

30 September, 2008 at 9:41 am
Terence Tao

Dear Pedro,

Thanks for the comments! I agree that the recent p
Klainerman-Rodnianski and others in establishing gauge-invariant analytic
physical (or configuration) space is very encouraging. There has also been s
progress in finding gauge-invariant substitutes for some tools that used to r
on frequency space, for instance using geometric heat flows as a substitute
Littlewood-Paley theory, or the spectral theory of the Laplacian as a substitu
Fourier transform. And of course we have microlocal analysis, which is alre
be invariant under canonical transformations and so has a good chance of h
reasonable gauge-invariance properties also. But the one thing we are still
have a gauge-invariant substitute for finer-scale frequency analysis, which i
coarse as Littlewood-Paley theory or as restricted to high-frequency or sem
limits as microlocal analysis. In particular, a key thing one wants to do with
separate them into pieces depending on their direction (or momentum); I d
a way to do this other than by invoking the Fourier transform (or related tra
such as Hilbert transforms, Riesz transforms, or Radon transforms) to work
frequency space (or momentum space), and this breaks all the gauge invari
are a few isolated papers that attempt to perform momentum decompositio
physical space means (e.g. by using various spacetime cutoffs) but progress
rather tentative.

At present, I am agnostic on which of these three general approaches (work
artificial (but analytically convenient) gauge, working with a "geometrically
gauge, or working in a gauge-invariant context) is "best" for these sorts of p
my guess is that we will need all three types of approaches, and be able to s
from one to the other when necessary.

Reply

1 October, 2008 at 7:16 am
Muhammad Alkarouri

By a purely lucky coincidence, I was just yesterday
at the concept of gauge yesterday, and I find this br
explanation.
As an aside, I would like to thank you (Prof. Tao) for the whole blog, and to

would you expect to publish the blog book you are preparing?

Back to the topic, in Wikipedia they are explaining a connection between a corresponding psudo-norm. Can you shed a little light on that please?
And does it make sense to connect the gauge explanation here with your id geometry of interpreting a linear transformation as a multidimensional gen a ratio?

Many thanks in advance,

Muhammad Alkarouri

Reply

1 October, 2008 at 10:20 pm
Roland Bacher

I think there is a tiny missprint in Example 1: It shou $\mathbb Z/2\mathbb Z\ltimes S^1$ (In order to rem notation easily, my thesis adviser told me that the acting group opens its mouth and tries to swallow t acts upon.) Thanks for your blog and best wishes, Roland Bacher

Reply

2 October, 2008 at 9:47 am
Terence Tao

Dear Roland: thanks for the correction (and for the mnemonic!).

Dear Muhammad: The concept of a gauge function as used in convex geome distantly related to the concept of a gauge for a bundle, though perhaps the natural way to interpret the former as a special case of the latter.

Multidimensional linear coordinate systems, which are given by linear trans are indeed multidimensional generalisations of one-dimensional coordinate which can be viewed as ratios between some physical quantity (e.g. a unit l numerical quantity (e.g. the number 1). But this is of course a rather specia ratio. The more typical ratios in practice connects one physical quantity to a a speed 30 m/sec is a ratio between length and time, or equivalently a linea transformation from the one-dimensional vector space of time displacement dimensional vector space of spatial displacements. Multidimensional transf e.g. velocity (a linear transformation from the one-dimensional vector space displacements to the three-dimensional space of spatial displacements), or a magnetic field acting on a charge (a linear transformation from the three- space of velocities to the three-dimensional space of forces) can thus be vie multidimensional ratio, given not by a single number, but as a matrix of nur indexed by the various degrees of freedom for the input and output.

I just sent off the final galley proofs for my book to the AMS, and hopefully

be done before the end of the year (at which point I suppose I will start wor
next volume.)

Reply

4 October, 2008 at 11:41 am
Allen Knutson

Two other related topics:

1. Quivers. This theory basically concerns connectio bundles over (usually finite) directed graphs. One novel feature, over manif the dimension of the "bundle" may change over different points. The "conn usually called a representation of the quiver, is a choice of linear map for ea the graph. If one fixes a gauge, i.e. a basis for each vector space, then the t boring — the space of connections is itself a big vector space. The interest i gauge transformations, whose group is the product of the general linear gr vector spaces.

In the most basic case, there is one edge. Then the nullity plus rank theore there exists a gauge in which the linear map is especially simple. Gabriel's says that there are only discretely many gauge-equivalent classes iff the gra

2. Currency trading (e.g. this paper). Here the finite graph is a complete di set of currencies, and the linear maps (between all 1-d spaces; this is electromagnetism) are exchange rates. The curvature (magnetic field) mea possibility of arbitrage, and abitrageurs are charged particles. This paper is read.

Reply

5 October, 2008 at 6:25 am
John Sidles

In hopes of keeping this gauge-theory thread (gently) stimulated—because IMHO many more readers of this weblog have good ideas than are posting—perhaps st be interested to learn that Shannon's classic 1949 article *Communication i Presence of Noise* begins with the sentence "A method is developed for rep any communication system *geometrically*" (and Shannon's article is still we reading today).

In the subsequent six decades, the geometric point-of-view pioneered by Sh flowered … and so has our algebraic, informatic, and combinatoric underst name just a few other mathematical disciplines).

A nice thing about gauge formalisms is that they provide a natural meeting these mathematical points of view. The good news for younger mathematici scientists, and engineers, and even economists) is that we still have a long understanding how these threads are most naturally united.

Reply

6 October, 2008 at 12:10 pm
Muhammad Alkarouri
Thank you very much, Prof. Tao
Reply

23 December, 2008 at 4:22 pm
Cohomology for dynamical systems « What's new
[…] (In this context, the first cohomology becomes a quotient space rather than a group; see a
post interpreting these cocycles in the language of gauge theory.) It seems to me that in this c
Reply

28 December, 2008 at 10:28 pm
Tricks Wiki: Use basic examples to calibrate exponents « What's new
[…] automatic by working exclusively with gauge-invariant notation (see also my earlier post o
theory). Another important test case for Schrödinger equations is the high-frequency limit , cl
Reply

10 January, 2009 at 12:14 pm
245B, notes 3: L^p spaces « What's new
[…] structure), isometric (to preserve metric structure), etc. Besides giving us useful symmetr
the presence of such group actions allows one to apply the powerful techniques of representat
Reply

26 January, 2009 at 7:54 am
Michael Nielsen » Doing science online
[…] other posts, on topics like Perelman's proof of
conjecture, quantum chaos, and gauge theory. Mar
remarkable insights, often related to open researcl
they […]
Reply

12 June, 2009 at 9:51 pm
Matt Cargo

Hi Terence,
In https://terrytao.wordpress.com/2008/09/27/what-is-a /#comment-32716
you said
"And of course we have microlocal analysis, which is already set up to be in under canonical transformations and so has a good chance of having reasor invariance properties also. But the one thing we are still missing is to have invariant substitute for finer-scale frequency analysis, which is not as coars Littlewood-Paley theory or as restricted to high-frequency or semi-classical microlocal analysis."

I studied this precisely this subject for my thesis at UC Berkeley. See for ex paper http://arxiv.org/abs/math-ph/0506074 , in which I show that, in theory creation and annihilation operators can be constructed out of the Weyl sym quantum integrable system. These lead directly to higher order corrections Sommerfeld quantization rules. There was a rub, unfortunately, which invol freedom: A

Reply

12 June, 2009 at 10:02 pm
Matt Cargo

[ahem, accidental return. To continue,]
At lowest order, there is the freedom in choosing the c angle variables. This is unfortunate because these var appear in the expression (interestingly, involving a symplectic connection) f correction to the symbols of the creation/annihilation operators. The expres gauge invariant, but I was never able find a way to write it using only gauge independent quantities, that is, with only the action variables. The problem every order, and is in fact due the overall phase freedom in the quantum wa

Reply

19 October, 2009 at 4:58 pm
Grothendieck's definition of a group « What's new


[…] transport") the fibre at the initial point of to the fibre at the final point; see this previous b more discussion. Note that the identity property is redundant, being implied by the other three
Reply

21 October, 2009 at 1:49 pm
Will M Farr

Dr. Tao,
I realize that I'm coming to this post late in the game wanted to give a data point regarding your musing that, "It is definitely con instance, that a given gauge field equation is well-posed with one choice of ill-posed with another." This is certainly the case for Einstein's equation in relativity, and has been a problem that the numerical relativity community l

on extensively over the last few decades! Depending on the choice of gauge character of the equations can change completely—see Paschalidis, Khokhl Novikov, arXiv:gr-qc/0511075 and Paschalidis, arXiv:0704.2861 for some w classifying common first-order formulations of Einstein’s equation and cons formulations that are well-posed in any gauge.

Thanks for the post—I love the clarity of your style in general, and in this p really hit the ball out of the park.

Reply

---

29 January, 2010 at 4:54 pm
Episode 005: 12-Gauge Theory – OR – How the Delorean Can Save Economics | Math f
[…] Terrence Tao on Gauge Theory […]
Reply

---

23 February, 2010 at 11:36 pm
254A, Notes 6: Gaussian ensembles « What’s new
[…] One approach here would be to artificially “fix a gauge” and work on some slice of the par which is “transverse” to all the symmetries. With such an approach, one can use the classical variables formula. While this can certainly be done, we shall adopt a more “gauge-invariant” a carry the various invariances with us throughout the computation. (For a comparison of the tw see this previous blog post.) […]
Reply

---

28 February, 2010 at 4:59 pm
Year Twenty Eight « Sarosh Wahla
[…] (b) Learn about gauge theory; […]
Reply

---

10 July, 2010 at 1:32 pm
Cayley graphs and the geometry of groups « What’s new
[…] fibre subgraph. The notion of a splitting in group theory is analogous to the geometric not gauge. The existence of such a splitting or gauge, and the relationship between two such split
Reply

---

21 July, 2010 at 9:16 am
it begins « 逝去日子
[…] it begins □□□ ： Learning,counter — □□□□□□□□ @ 11:27 □
[…]
Reply

---

17 August, 2010 at 6:31 am
counters « 逝去日子
[…] .later gauge/integrability-V.E.Zakharov/Lars Onsager/dipole/di analysis/ill-posed problems/Pascal […]
Reply

---

8 December, 2010 at 6:57 pm
Neurociência e o Projeto Ersätz-Brain… « Ars Physica
[…] de redes neurais via sistemas dinâmicos, modelo de Potts e, por que não, teorias de gauge gauge?, Gauge theories (scholarpedia), Preparation for Gauge Theory e Gauge Theory (José [..
Reply

---

2 November, 2011 at 1:33 pm
Smolin (2011) | Research Notebook
[…] the remainder of this section, I use notes from Gan (1999), Singer (2001) and Tao (2008). Below I give a de fiber bundle: Definition (Fiber Bundle): A fiber bundle

Reply

12 December, 2011 at 4:59 pm
nlcatter
you started out fine and then diverged to uselessne

Reply

30 October, 2012 at 8:42 pm
artojh
Well written and explained, does the work of Ruđer J Bošković and his writing on relativity in his volumes in 1785 have a bearing on the understanding of his u co-ordinates well before the modern set of guage theories. Thanks Arto

Reply

29 December, 2012 at 1:04 pm
A mathematical formalisation of dimensional analysis « What's new

[…] time. (This is closely related to the concept of spending symmetry, which I discuss for insta (or in Section 2.1 of this […]

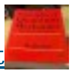
Reply

30 December, 2012 at 6:24 am
Anonymous
wow, much love and appreciation. This just made m smarter physics student. I could not begin to thank enough. It was clear concise and beautiful.

Reply

9 June, 2013 at 1:51 pm
Daniel Dobkin
Beautifully clear, with lovely simple examples to clarify v heck the obscure abstract definitions mean, in contrast t number of other posts and pages I've read on the topic. remembered my differential geometry from Misner, Thorne and Wheeler th back I would have actually followed the details! But now I know what a fibe and why you might need to transport it. Thanks.

Reply

27 October, 2013 at 3:14 pm
What is a gauge? | What about being a physicist

[…] What is a gauge?. […]

Reply

1 December, 2013 at 1:50 am
Simon Crase
Thank you for a clear exposition. I had been working way the The Road to Reality until I hit a brick wall w Fibre Bundles & Gauge Theory. Thanks to you I have though. And I'm making sense of some of the scatterd bricks, too…;-)

Reply

13 February, 2014 at 8:24 pm
What is the Significance of Lie Groups $SO(3)$ and $SU(2)$ to Particle Physics? | We

Animals
[…] Terrence Tao's blog "What is a gauge?" […]
Reply

28 March, 2014 at 9:02 am
Quora
What is Gauge Theory (intuitively)?
Here's a great answer from Terence Tao: https://terrytao.wordpress
/27/what-is-a-gauge/
Reply

15 July, 2014 at 4:28 am
Varun

Dear Prof. Tao,

Thank you very much for such a clear exposition on gau
bachelor's student, recently been trying to understand Donaldson and Kron
exposition The Geometry of Four Manifolds.

I have a somewhat naive question. I came across the following comment: "
equation is non-linear, but more to the point it is not elliptic, i.e, the highest
d^+ is not elliptic. This is clear from abstract grounds from the invariance of
under gauge transformations."

How is the gauge group responsible for the ellipticity of the operator?

Varun

Reply

15 July, 2014 at 12:07 pm
Terence Tao

Elliptic operators should have uniqueness of the Dirich
problem, but gauge symmetry implies lack of uniquene
one can look at how elliptic regularity is incompatible v
symmetry.)

Reply

29 August, 2014 at 3:55 am
[转载]What's a Gauge? | 小小泪
[…] Terence Tao's blog: What's a Gauge. "Gauge theory"
has connotations of being a fearsomely complicated part
Reply

28 December, 2014 at 6:20 am
Terry Tao's "what is a gauge?" | The Daily Pochemuchka
[…] the introduction to the blog
[…]
Reply

26 March, 2016 at 4:28 am
Nistwo Rai

Can anyone help me.
What do you mean by saying half maximally gauged a
maximally gauged????

Reply

13 June, 2016 at 12:34 am
the displacement vector arises due to gauge transform

imran khan

taken a time dependent, any one tell me, can we take it as a function of “r” i-e B(r) instead of B(t).

Reply

---

25 July, 2016 at 10:39 am
Omissions in Mathematics Education: Gauge Integration - Comments | Page 2 | Physi The Fusion of Science and Community
[…] concept of a "gauge" in physics isn't very intuitive. There are explanations such as https://terrytao.wordpress.com/2008/09/27/what-is-a-gauge/ , but I need someone to explain th Stephen Tashi, Jul 25, 2016 at […]
Reply

---

27 December, 2016 at 10:28 pm
Terry Bollinger

Very minor erratum: Coordinate Systems, #8 says $x^2 + y^2 + z^2$; perhaps $x^2 + y^2 + z^2 = 1$ was intended?

[Corrected, thanks – T.]

Reply

---

3 April, 2017 at 7:08 am
Symmetry → connection field → force of nature – Physics says what?

[…] https://terrytao. /2008/09/27/what-is technical explanatio

Reply

---

26 April, 2017 at 7:04 am
JR

I have one objection to this otherwise brilliant article. I do not think it is meaningful to say that a gauge transformation is necessarily a coordinate transformati all physical theories should be invariant under arbitrary coordinate transfor This is the reason why the group of diffeomorphisms of spacetime cannot b gauge group of any physical theory unless of course, you consider active diffeomorphisms. In the physics community, the term “gauge theory” is rese much more restrictive class of physical theories.

Reply

---

6 June, 2017 at 1:01 pm
Anonymous

Tao….the most professional young mathematician

Reply

---

22 August, 2017 at 4:36 pm
An addendum to “arbitrage, amplification, and the tensor power trick” | What's new
[…] to proving inequalities such as (1). There is a complementary approach, discussed for inst previous post, which is to spend the symmetry to place the variable “without loss of generality
Reply