

Chapter 2: A few introductory remarks on the relations between topology, manifolds, and group theory, with some applications to physics

ABSTRACT: December 22, 2014

Contents

1. Introduction	5
2. Some motivating examples	5
2.1 Some formulae from classical electrodynamics	5
2.2 The Gauss linking number of two closed curves in three dimensional space	7
2.2.1 Relation to the Hopf invariant	10
2.3 Angular momentum of a pair of dyons	12
2.4 Dirac quantization	13
2.5 Quantum mechanics of a charged particle moving in a classical electromagnetic field	14
2.6 A charged particle in a flat gauge field	18
2.7 Dirac's argument for quantization	19
2.8 Anyons in 2+1 dimensions	20
2.9 The moral of the examples	23
3. Some basic definitions in topology	24
3.1 Definition of a topology	24
3.2 Continuity and convergence	26
3.3 Homeomorphism	27
3.4 Compactness	29
3.5 Connectedness	30
4. Topologies on spaces of functions	31
5. Constructing topological spaces using gluing	35
6. Manifolds	37
6.1 Basic Definitions	38
6.2 Examples	43
6.3 Tangent and cotangent space	47
6.3.1 Definition of $T_p M$ using directional derivatives	47
6.3.2 Definition of $T_p M$ in terms of derivations	52
6.3.3 First definition of $T_p^* M$	54
6.3.4 Algebraic definition of $T_p^* M$	54
6.3.5 The differential of a map	56
6.4 Orientability	57
6.5 Tangent and cotangent bundles	59
6.5.1 Example 1: TS^2 and T^*S^2	62
6.6 Definitions: Submersion, immersion, embedding, critical and regular points and values	63

6.7	Whitney embedding theorem	66
6.8	Local form of maps between manifolds and of submanifolds	67
6.8.1	Local form of a differentiable map between manifolds	67
6.8.2	Local form of a submanifold	70
6.8.3	Defining submanifolds by equations	71
6.9	Lie groups	76
6.9.1	Lie algebras of Lie groups	78
6.9.2	Remarks on the classification of Lie groups	82
6.10	Transversality	84
6.10.1	Relative Transversality	88
6.11	Intersection Numbers	88
6.11.1	The Whitney disk trick	90
6.11.2	Intersection pairing and homology theory	92
6.12	Linking	93
6.13	Introduction to singularity theory	95
6.13.1	Some motivations	96
6.13.2	Canonical forms of functions	97
6.13.3	Germes, Jets, and Unfoldings	101
6.13.4	Some Examples	102
6.13.5	Maps between manifolds	104
6.13.6	Complex singularities	105
6.13.7	Some sources	107
6.14	Digression: Classification of manifolds	107
6.14.1	Three categories of manifolds	108
6.14.2	Four dimensions	110
6.14.3	The Generalized Poincaré conjecture	114
6.14.4	Sources	115
7.	Transformation Groups, Group Actions, and Orbits	116
7.1	Definitions and the stabilizer-orbit theorem	116
7.1.1	The stabilizer-orbit theorem	119
7.2	First examples	120
7.3	Action of a topological group on a topological space	124
7.4	Left and right group actions of G on itself	129
7.5	Induced group actions on function spaces	130
7.5.1	Application: Functions on groups	131
7.6	An Example of Orbits in physics: Orbits of the Lorentz group and relativistic wave equations	137
7.6.1	The case of $1 + 1$ dimensions	137
7.6.2	Orbits, Representations, and Differential Equations	139
7.6.3	The massless case in $1 + 1$ dimensions	140
7.6.4	The case of d dimensions, $d > 2$	142
7.7	Spaces of orbits	144

7.7.1	Simple examples	146
7.7.2	Fundamental domains	147
7.7.3	Algebras and double cosets	152
7.7.4	Orbifolds	152
7.7.5	Examples of quotients which are not manifolds	153
7.7.6	When is the quotient of a manifold by an equivalence relation another manifold?	157
8.	Homogeneous spaces of Lie groups	158
8.1	Grassmannians	162
8.1.1	Homogeneous spaces	162
8.1.2	Coordinates and coordinate patches	164
8.1.3	Orthonormal bases	166
8.1.4	Schubert cells	168
9.	Bundle Basics	170
10.	The classification of compact two-dimensional surfaces	171
11.	Homotopy of maps and spaces	182
11.1	Homotopy of maps	182
11.2	Homotopy of maps of pairs	185
11.2.1	Example: Homotopy of curves	186
12.	Homotopy groups	187
12.1	π_0	187
12.2	The fundamental group: π_1	187
12.2.1	Remark on winding number	191
12.2.2	Surface groups	192
12.2.3	Braid groups	194
12.2.4	Digression: A_∞ spaces	196
12.3	Higher homotopy groups	203
12.4	Homotopy groups and homotopy equivalence	206
12.4.1	Homotopy Invariants of maps between spaces	207
12.5	Homotopy and its relation to loop space	207
13.	Fibrations and covering spaces	208
13.1	The lifting problem	208
13.2	Homotopy lifting property	209
13.3	The long exact sequence of homotopy groups for a fibration	213
13.4	Covering spaces	221
13.5	Path lifting, connections, monodromy, and differential equations	223
13.5.1	Interlude: Ordinary differential equations	224
13.6	Solution of the lifting problem for covering spaces	233

13.7	The universal cover	233
13.8	The Galois correspondence between covers of X and subgroups of $\pi_1(X)$	239
13.8.1	Galois correspondence and normal subgroups	242
13.9	Coverings and principal bundles with discrete structure group	244
13.10	Branched covers and multi-valued functions	246
13.10.1	Example: Hyperelliptic curves	247
13.10.2	Riemann-Hurwitz formula	249
14.	CW Complexes	249
14.1	The Euler character	252
15.	Bordism and Cobordism	252
16.	Counting solutions of an equation: The degree of a map	252
16.1	Intersection interpretation	255
16.2	The degree for proper maps between manifolds	256
16.3	Examples	257
16.4	Computing $\pi_n(S^n)$	258
16.5	The degree as a “topological field theory integral”	259
17.	Overview of the uses of topology in field theory	259
17.1	Digression: Physics and the classification of manifolds	260
18.	Solitons and soliton sectors	260
18.1	Soliton sectors	260
18.2	A simple motivating example: Solitons in the theory of a scalar field in 1+1 dimensions	261
18.3	Landau-Ginzburg solitons in 1+1 dimensions	266
18.4	Minkowskian spacetime of dimension greater than two	270
18.5	Solitons in spontaneously broken gauge theories	271
18.6	The general field theory of scalar fields: The nonlinear sigma model	273
19.	“Instanton” sectors	275
19.1	Fieldspace topology from boundary conditions	275
19.2	A charged particle on a ring around a solenoid, at finite temperature	276
19.3	Worldsheet instantons in string theory	281
19.4	MORE EXAMPLES	282
20.	Sources	282

1. Introduction

2. Some motivating examples

Topology plays a very important role in modern physical mathematics. Unfortunately, many of the most exciting applications require some background in quantum field theory and string theory. The purpose of the simple examples in this section is to demonstrate how simple computations in electrodynamics and quantum mechanics can lead to some interesting physical quantities that exhibit interesting topological invariance, or which have interesting connections to nontrivial constructions in topology.

2.1 Some formulae from classical electrodynamics

We are going to do some computations in electrodynamics, so let us set out a few conventions and summarize the basic formulae of electrodynamics here.

The fieldstrength $F_{\mu\nu}$ is best regarded as a 2-form

$$F = \frac{1}{2!} F_{\mu\nu} dx^\mu dx^\nu \quad (2.1)$$

If we have a splitting into space and time we can define electric and magnetic fields by setting $x^0 = ct$, t is time and

$$F = E \wedge dx^0 + B \quad (2.2)$$

so $F_{i0} = \vec{E}_i$.

The action for a Maxwell field in d -dimensional Minkowski space is:

$$S = \frac{1}{8\pi} \int (\vec{E}^2 - \vec{B}^2) d^d x = -\frac{1}{16\pi} \int F_{\mu\nu} F^{\mu\nu} d^d x \quad (2.3)$$

A word about units: Our conventions are close to the cgs, or Gaussian, conventions. Clearly, the electric and magnetic fields have the same units $(ML^{2-d}T^{-1})^{1/2}$.¹ The difference with cgs/Gaussian units as used, for example, in J.D. Jackson, *Classical Electrodynamics*, 2nd edition, is that the fields (and charges) are related by a multiplicative factor of $c^{1/2}$. So $\vec{E}^{here} = c^{-1/2} \vec{E}^{Jackson}$, etc.

The energy momentum tensor is

$$T_{\mu\nu} = \frac{1}{4\pi} \left(F_{\mu\lambda} F_\nu{}^\lambda - \frac{1}{4} g_{\mu\nu} F_{\rho\lambda} F^{\rho\lambda} \right) \quad (2.4)$$

so in particular the energy density is $T_{00} = \frac{1}{8\pi} (\vec{E}^2 + \vec{B}^2)$ while the momentum density is $T_{0i} = \frac{1}{4\pi} F_{0j} F_i{}^j$. In $\mathbb{M}^{1,3}$ this can be written $T_{0i} = -\frac{1}{4\pi} \epsilon_{ijk} E_j B_k$. The integral over space gives the energy in the field:

$$E = cP_0 = c \int T_{00} d^{d-1} \vec{x} \quad (2.5)$$

¹We indicate units with $M^x L^y T^z$ for mass (M), length (L) and time (T) units. If we choose units where $c = 1$ then we identify L and T . If we further choose units where $\hbar = 1$ we then identify $M = 1/L = 1/T$.

and the momentum:

$$P_i = \int T_{0i} d^{d-1} \vec{x}. \quad (2.6)$$

In any number of dimensions the rotations about a point in space $\vec{a} \in \mathbb{R}^{d-1}$ are generated by the rotations in the $x^i x^j$ -planes. For the electromagnetic field, the angular momentum for rotation in the $x^i x^j$ plane is:

$$J_{ij} = \frac{1}{4\pi} \int d^{d-1} \vec{x} ((x-a)_i T_{0j} - (x-a)_j T_{0i}) \quad (2.7)$$

When coupling to a source with current j_μ we take the action:

$$S = -\frac{1}{16\pi} \int F_{\mu\nu} F^{\mu\nu} d^d x + \int j_\mu A^\mu \quad (2.8)$$

These lead to Maxwell's equations (viewing j_μ as a 1-form, which is the correct point of view):

$$\begin{aligned} d * F &= 4\pi j \\ dF &= 0 \end{aligned} \quad (2.9)$$

In particular, for a particle of charge q moving on a worldline $x^\mu(s)$ we get

$$S = -\frac{1}{16\pi} \int F_{\mu\nu} F^{\mu\nu} d^d x + \int q A_\mu(x(s)) \frac{dx^\mu}{ds} ds \quad (2.10)$$

The electric field produced by a static electric charge q at position $\vec{x} = \vec{R}$ satisfies

$$\nabla_i \vec{E}_i = 4\pi q \delta^{(d-1)}(\vec{x} - \vec{R}) \quad (2.11)$$

and is given, explicitly, by

$$\vec{E} = \kappa_d q \frac{\vec{x} - \vec{R}}{|\vec{x} - \vec{R}|^{d-1}} \quad (2.12)$$

where

$$\kappa_d = \frac{4\pi}{\text{vol}(S^{d-2})} \quad (2.13)$$

and $\text{vol}(S^{d-2})$ is the volume of the unit sphere in the metric induced from Euclidean space. In general

$$\text{vol}(S^n) = \frac{2\pi^{(n+1)/2}}{\Gamma(\frac{n+1}{2})} \quad (2.14)$$

Exercise *Units of charge*

Show that in our conventions a charge q has units of $(ML^{d-2}T^{-1})^{1/2}$.

Exercise

Compute $\eta^{\mu\nu}T_{\mu\nu}$.

Observe that it vanishes for $d = 4$. This is very significant. It means that Maxwell theory is a classical conformal field theory in $3 + 1$ dimensions.

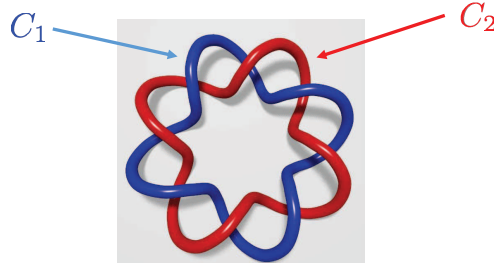


Figure 1: Two wires link. Figure taken from Wikipedia article on linking number.

2.2 The Gauss linking number of two closed curves in three dimensional space

Let us consider the following, somewhat odd, construction in magnetostatics, one which, however, was already considered by J.C. Maxwell himself in his great book.²

The problem is this: *Consider a closed oriented loop $C \subset \mathbb{R}^3$ carrying a current I . Next, consider a second oriented loop $C' \subset \mathbb{R}^3$ as in Figure 1. What is the work done by the magnetic field when transporting a magnetic pole of unit charge around the curve C' ?*

We claim that the result is given by a *topological invariant*, the *Gauss linking number*, which measures the amount by which the loops C and C' are linked.

We begin with one of Maxwell's equations allowing us to derive the magnetic field from a current source:

$$\vec{\nabla} \times \vec{B} = \vec{J} \quad (2.15)$$

where $\vec{B}$ is the magnetic field and $\vec{J}$ is the current density.

The current density is:

$$J^k(\vec{x}) = I \oint_C \frac{dx^k(s)}{ds} \delta^{(3)}(\vec{x} - \vec{x}(s)) \quad (2.16)$$

where $\vec{x}(s)$ describes the loop C and s is a monotonically increasing parameter along the loop.

The work done by the magnetic field on a magnetic pole transported around C' is:

$$\oint \vec{B} \cdot d\vec{\ell} \quad (2.17)$$

²J.C. Maxwell, *A Treatise on Electricity and Magnetism*, Section 419. Dover 1954, vol. 2, p.43

We claim this is given by

$$\oint \vec{B} \cdot d\vec{\ell} = IL(C, C') \quad (2.18)$$

where

$$L(C_1, C_2) := -\frac{1}{4\pi} \int_{C_1} \int_{C_2} \frac{(\vec{x}_1 - \vec{x}_2) \cdot (\frac{d\vec{x}_1}{ds_1} \times \frac{d\vec{x}_2}{ds_2})}{|\vec{x}_1 - \vec{x}_2|^3} ds_1 ds_2 \quad (2.19)$$

where $\vec{x}_1 := \vec{x}_1(s_1)$ describes the loop C_1 and $\vec{x}_2 := \vec{x}_2(s_2)$ describes the loop C_2 . Note that $L(C_1, C_2) = L(C_2, C_1)$.

In order to prove (2.19) note that we can solve for B using the Biot-Savart law:

$$\vec{B}(\vec{r}_1) = -\frac{I}{4\pi} \vec{\nabla}_1 \times \int_C \frac{d\vec{r}_2}{|\vec{r}_1 - \vec{r}_2|} := -\frac{I}{4\pi} \vec{\nabla}_1 \times \int_0^1 \frac{1}{|\vec{r}_1 - \vec{r}_2(t)|} \frac{d\vec{r}_2(t)}{dt} dt \quad (2.20)$$

Remark: The integral formula for $L(C_1, C_2)$ was discovered by Gauss in 1833. It has some similarities with, but is really quite different from, the Neumann formula for the mutual inductance of two current loops.

We claim that $L(C, C')$ is in fact an *integer* which measures the linking of C and C' . In order to show that $L(C, C')$ is indeed an integer it is useful (but not strictly necessary) to use differential forms.

Conceptually, the magnetic field should be combined together with the electric field to make a two-form F on Minkowski spacetime $\mathbb{M}^{1,3}$, and then (2.15) is a special case of $d * F = j$.

However, in magnetostatics, if we do not worry about orientation-reversing spacetime transformations, we can think of $\vec{B}$ as defining a one-form on $\mathbb{R}^3$, using $B = B^i dx^i$. Similarly, we can think of the current $\vec{J}$ as defining a two-form $J = \frac{1}{2} \epsilon_{ijk} J^i dx^j dx^k$. We are using here the Hodge duality between one-forms and two-forms in $\mathbb{R}^3$ and the equivalence of vectors and one-forms from a Euclidean metric.

In any case, identifying B with a one-form and J with a two-form, equation (2.15) is equivalent to:

$$dB = J \quad (2.21)$$

and the work done is just $\int_{C'} B$.

Now put $I = 1$ and let D' be an oriented disk spanning C' as in Figure 2. Then we evaluate the total current flowing through D' :

$$\begin{aligned} \oint_{C'} B &= \int_{D'} dB = \int_{D'} J \\ &= \int_{D'} d\xi^\alpha \wedge d\xi^\beta \frac{\partial x^m}{\partial \xi^\alpha} \frac{\partial x^n}{\partial \xi^\beta} \frac{1}{2} \epsilon_{mnj} \oint_C \frac{dx^j(t)}{dt} \delta^{(3)}(\vec{x}(\xi) - \vec{x}(t)) \end{aligned} \quad (2.22)$$

where ξ^α are some coordinates on D' . It is easy to see that in the last expression each transverse *intersection* of D' with C contributes ± 1 according to orientation: The orientation of C' induces one on D' , and C is oriented. This oriented intersection number is

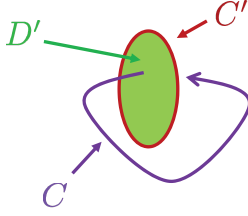


Figure 2: If we fill in C' with a disk D' then we can show that $L(C, C')$ counts signed intersections of C' with D' , and is therefore an integer.

one of the definitions of the linking number. From this interpretation $L(C, C')$ is clearly invariant under continuous deformation of D' or C or C' , so long as C and C' do not cross.

Now that $L(C, C')$ is a *continuous* function of the locations of C and C' . On the other hand it is an *integer*. Therefore, it is a *topological invariant*. Note that this topological invariant can change, if we allow C and C' to cross. When C and C' cross the formula (2.19) becomes ill-defined, and then the integer can jump.

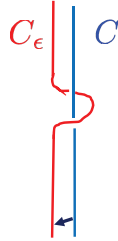


Figure 3: Displacing C infinitesimally to C_ϵ in the normal direction might lead to nontrivial self-linking because the normal vector might twist around. For this reason $L(C, C)$ is ill-defined.

Just as self-inductance is rather more subtle than mutual inductance, self-linking numbers are a good deal more subtle than mutual linking numbers. Let us continue to take $I = 1$, and let $B^{(C)}$ be the resulting magnetic field from a closed loop C . The two-form J is an example of what is known as a *representative of a Poincaré dual to C* which means, roughly speaking, that integrals with J localize onto integrals on C . In particular, we could

try to define the self-linking number by

$$\begin{aligned}
L(C, C) &\stackrel{?}{=} \int_C B^{(C)} = \int_C B_i^{(C)}(\vec{x}(t)) \frac{dx^i(t)}{dt} \\
&= \int_{\mathbb{R}^3} B^{(C)} \wedge J \\
&= \int_{\mathbb{R}^3} B^{(C)} \wedge dB^{(C)}
\end{aligned} \tag{2.23}$$

This last integral is known as the *Chern-Simons invariant* of the one-form $B^{(C)}$ on $\mathbb{R}^3$.

The Chern-Simons invariant is well-defined for a *smooth* one-form, but, due to the current source, $B^{(C)}$ has singularities on C . Consequently, this self-linking number is ill-defined. That makes perfectly good sense, as explained in Figure 3. To define $L(C, C)$ one must displace C infinitesimally in a normal direction and evaluate the mutual linking number of C and its displaced version. but this is clearly ill-defined because C could link around itself several times. In Chern-Simons theory this is known as the *framing anomaly*.

Exercise *Explicit verification of topological invariance*

Show that $I(C_1, C_2)$ is invariant under small deformations of $\vec{x}_1(t)$ by an explicit variation of the formula (2.19).

Exercise

Define the 2-form on $\mathbb{R}^3 - \{0\}$

$$\omega(x) := \frac{1}{8\pi} \frac{\epsilon_{ijk} x^i dx^j dx^k}{|x|^3} \tag{2.24}$$

a.) Show that, when restricted to the unit sphere $S^2 \subset \mathbb{R}^3 - \{0\}$ the form ω restricts to the standard volume form with unit volume.

b.) Show that the Gauss linking number can be expressed elegantly as:

$$L(C_1, C_2) = \int_{C_1} \int_{C_2} \omega(x_1 - x_2) \tag{2.25}$$

2.2.1 Relation to the Hopf invariant

The above innocent observation is actually related to a deep mathematical construction known as the *Hopf fibration*, which is a particularly beautiful map $p_H : S^3 \rightarrow S^2$. We will describe it later when we discuss the representation theory of $SU(2)$. The following remarks use much material that you are not expected to know.

The regularized self-linking number, which is the regularized Chern-Simons invariant, is related to something known as the Hopf invariant of a map $f : S^3 \rightarrow S^2$, denoted $H(f)$. A beautiful aspect of the Hopf invariant is that it measures the linking number between the preimages of regular points $f^{-1}(a)$ and $f^{-1}(b)$. In particular, if f is the projection p_H of the Hopf fibration then any two fibers will be linked since $H(p_H) = 1$.

We connect the Hopf invariant and linking numbers of the fibers to the above exercise in electromagnetism as follows. We introduce stereographic projection from a pole, call it the north pole. So we have

$$\begin{array}{ccc} S^3 - \{(1, 0, 0, 0)\} & \xrightarrow{p_N} & \mathbb{R}^3 \\ p_H \downarrow & & \\ S^2 & & \end{array} \quad (2.26)$$

Now we can choose a representative of a generator of $H^2(S^2; \mathbb{Z})$, call it ω , which is a bump form centered on $\hat{n}_0 \in S^2$, and in the limit as the support goes to zero $\omega \rightarrow \delta^{(2)}(\hat{n} - \hat{n}_0)d^2\hat{n}$. Note however, that the bump form is perfectly smooth. Then $p_H^*(\omega)$ is supported on the fiber above $\hat{n}_0$ and in the limit that the support of the bump form shrinks to $\hat{n}_0$

$$J = (p_N^{-1})^* p_H^* \omega = J^{(C)} \quad (2.27)$$

is the current on a wire C where C is a closed loop in $\mathbb{R}^3$ which is the stereographic projection of the Hopf fiber over $\hat{n}_0$. Since $H(f) = 1$ for the Hopf fibration we see that, with the regularization provided by the smooth bump form, the singular self-linking number becomes well-defined and indeed must be $L(C, C) = 1$. This is of course true for any choice of $\hat{n}_0$. Note further that

$$p_N^* B^{(C)} = \Theta \quad (2.28)$$

relates the corresponding “magnetic field” to the connection one form Θ on the total space of the principal $U(1)$ bundle $\pi_H : S^3 \rightarrow S^2$.

Now, let us consider two points $\hat{n}_1, \hat{n}_2 \in S^2$ and let us represent the generator of $H^2(S^2)$ by

$$\omega_s = s\delta^{(2)}(\hat{n} - \hat{n}_1)d^2\hat{n} + (1-s)\delta^{(2)}(\hat{n} - \hat{n}_2)d^2\hat{n} \quad (2.29)$$

where s is any real number, and again we regularize the delta-functions by smoothing them out to bump forms. The B -field associated with this representative of a generator of $H^2(S^2)$ will be

$$B = sB^{(C_1)} + (1-s)B^{(C_2)} \quad (2.30)$$

where C_1, C_2 are the stereographic projections of the fibers above $\hat{n}_1, \hat{n}_2$, respectively. But now we have

$$\begin{aligned} 1 &= \int \Theta d\Theta = s^2 L(C_1, C_1) + 2s(1-s)L(C_1, C_2) + (1-s)^2 L(C_2, C_2) \\ &= 1 - 2s(1-s) + 2s(1-s)L(C_1, C_2) \end{aligned} \quad (2.31)$$

As indicated, the integral on the LHS is independent of representative ω and equal to 1. Therefore, $L(C_1, C_2) = 1$ for the images of any two fibers!!

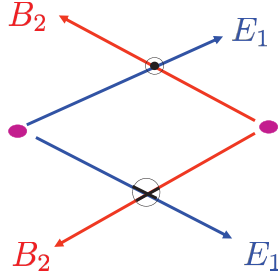


Figure 4: The angular momentum in the electromagnetic field of a pair of dyons. Illustrated here is one contribution to $\vec{E} \times \vec{B}$ from the electric field of particle 1 and the magnetic field of particle 2. Of course there is another contribution, not shown, from the magnetic field of particle 1 and the electric field of particle 2.

2.3 Angular momentum of a pair of dyons

Another simple computation in electromagnetism leads to an intriguing topologically invariant quantity.

Although particles with magnetic charge, so-called *magnetic monopoles*, have never been observed experimentally, one can investigate them theoretically, and they lead to some of the deepest modern mathematics.

A point particle in $\mathbb{R}^3$ at a location $\vec{R}$ of electric and magnetic charge (q, g) would produce an electric field

$$\vec{E} = q \frac{\vec{x} - \vec{R}}{|\vec{x} - \vec{R}|^3} \quad (2.32)$$

as well as a magnetic field ³

$$\vec{B} = g \frac{\vec{x} - \vec{R}}{|\vec{x} - \vec{R}|^3} \quad (2.33)$$

A particle with charges $q = 0$ and $g \neq 0$ is called a *magnetic monopole*. A particle with $g \neq 0$ where the charge q is also not necessarily zero is called a *dyon*.

Now consider the angular momentum of the electromagnetic field in the presence of a pair of dyons where we choose the origin to be at the midpoint of the segment separating the dyons. Specializing the formula (2.7) for the angular momentum to $d = 3 + 1$ dimensions the generator of rotations around a point $\vec{a}$ is:

$$\vec{L} = \frac{1}{4\pi} \int_{\mathbb{R}^3} (\vec{x} - \vec{a}) \times (\vec{E} \times \vec{B}) d^3\vec{x} \quad (2.34)$$

A little calculation, left as an exercise, shows that

$$\vec{L} = \hat{r}_{12} (g_1 q_2 - g_2 q_1) \quad (2.35)$$

³We use g and not m to avoid confusion with a mass.

where $\hat{r}_{12}$ is the unit vector pointing from particle 2 towards particle 1.

The result is somewhat remarkable: *It is independent of the distance between the dyons.* In this sense it is “topological.” It also leads to a very deep connection of topology with physics through Dirac quantization.

Exercise

a.) Derive equation (2.35).

b.) In general, due to the factor of $\vec{x}$ in the formula for the angular momentum, the angular momentum of an electromagnetic field depends on the choice of origin of spacetime. Show that for the case of two dyons the result is independent of the choice of origin of spacetime.

2.4 Dirac quantization

The computation of (2.35) was a classical computation, but it has extremely interesting implications when combined with quantum mechanics. Eventually, we will want to quantize the electromagnetic field as well. Then $\vec{L}/\hbar$ becomes a generator of the Lie algebra $so(3) \cong su(2)$ of the rotation group. We will study this Lie algebra later and see that the generators always have eigenvalues which are integral or half-integral. That is, of the form $n/2$ where $n \in \mathbb{Z}$. This leads to the *Dirac quantization law* (sometimes called the *Dirac-Schwinger-Zwanziger quantization law*):

$$g_1 q_2 - g_2 q_1 = \frac{1}{2} n \hbar \quad (2.36)$$

Remarks:

1. Among other things, this means that the existence of just one dyon somewhere in the Universe would explain the fact that electric charge is quantized.
2. Since the photon is a boson one might think that the energy n in (2.36) must be even. In fact, as an alternative argument below shows, this is not the case, and the field of a pair of dyons can very well have half-integral spin!
3. In this remark let us switch to more standard units for charge by scaling charges by a factor of $c^{-1/2}$. Then the formula for the fine structure constant is $e^2/\hbar c$. It is dimensionless and (at long distances) is approximately $1/137$. Similarly, the magnetic fine structure constant $g^2/(\hbar c)$ is dimensionless. The Dirac quantization condition implies

$$\frac{g^2}{\hbar c} \frac{e^2}{\hbar c} = \frac{n^2}{4} \quad n \in \mathbb{Z} \quad (2.37)$$

and hence in nature the magnetic fine structure constant $\frac{g^2}{\hbar c} \sim \frac{137}{4} n^2$ would be rather large. The energy loss of a relativistic monopole passing through matter is similar to that of a relativistic heavy nucleus of $Z = 137n/2$. It would have very dramatic effects and would be hard to miss. See Jackson, p. 253 for more details.

4. The equations with electric and magnetic sources are mathematically consistent, and perfectly natural in view of the electric-magnetic duality of Maxwell's equations. There is no fundamental reason for electric charge quantization in quantum electrodynamics, or even in the $SU(3) \times SU(2) \times U(1)$ standard model. Moreover, in nonabelian Yang-Mills-Higgs theories with suitable patterns of spontaneous symmetry breaking, a generic prediction is the existence of magnetic monopoles. A standard physical context for such theories are Grand Unified Theories in which case the monopole mass turns out to be of order the GUT scale $\sim 10^{15} GeV$. Moreover, in theoretical arguments, magnetic monopoles play a prominent role in explanations of the phenomenon of confinement in the strong interactions. One of the remarkable and deep aspects of Nature is that – to amazing accuracy – there are no observed magnetic sources in Nature. This is not for want of trying to find them. Searches have included:

1. Searches in rocks, moonrocks etc. for traces of ionization.
2. Induction experiments. Except for a notorious event on Valentine's day, 1982, there have been no observations, despite an extensive effort.
3. The existence of long-range galactic magnetic fields provides stringent bounds on the density of monopoles.
4. According to the PDG, the flux of massive magnetic monopoles with $1.1 \times 10^{-4} < \beta < 0.1$ is

$$flux < 1.0 \times 10^{-15} cm^{-2} sr^{-1} s^{-1}$$

The reason for the absence of monopoles in nature is a deep mystery.

5. In spite of the negative experimental situation, Dirac quantization has played a very important role in the development of physical mathematics.

2.5 Quantum mechanics of a charged particle moving in a classical electromagnetic field

The quantum wavefunction of the charged particle traveling on a path γ in a spacetime $\mathcal{S}$ has a multiplicative contribution to its phase which is of the form

$$\exp[i \frac{q}{\hbar} \int_{\gamma} A] \tag{2.38}$$

where $A = A_{\mu} dx^{\mu}$ is the “vector potential” of the electromagnetic field. (Locally, A is a one-form; better, it is a connection on a principal $U(1)$ line bundle over $\mathcal{S}$.)

Written out in detail: We describe the worldline by a map: $x : \mathcal{D} \rightarrow \mathcal{S}$ where the domain $\mathcal{D}$ is a subset of the real line. $\mathcal{D}$ could be the entire real line, or just a finite interval $\mathcal{D} = [s_i, s_f]$ or a half-line. When talking about thermodynamics or partition functions we take $\mathcal{D}$ to be a circle.

If we choose coordinates x^{μ} on $\mathcal{S}$ (with $\mu = 0, \dots, d-1$, where d is the dimension of spacetime), then the trajectory is described locally by functions $x^{\mu}(s)$, where $s \in \mathbb{R}$ is a

parameter. Then:

$$\exp[i\frac{q}{\hbar} \int_{\gamma} A] = \exp[i\frac{q}{\hbar} \int_{\mathcal{D}} A_{\mu}(x(s)) \frac{dx^{\mu}}{ds} ds] \quad (2.39)$$

Specializing to d -dimensional Minkowski space $\mathcal{S}_4 = \mathbb{M}^{1,d-1}$ with signature $(-1, +1^{d-1})$ the wavefunction of a charged particle of mass m and charge q moving in an electromagnetic field described by $A = A_{\mu} dx^{\mu}$ can be computed by the path integral over the space of all (sufficiently differentiable) functions

$$Map(\mathcal{D}, \mathbb{M}^{1,d-1}) = \{x : \mathcal{D} \rightarrow \mathbb{M}^{1,d-1}\} \quad (2.40)$$

with weighting (here s has units of time):

$$\int [dx^{\mu}(s)] \exp \left[\frac{i}{\hbar} \int_{\mathcal{D}} \left\{ mc \sqrt{-\frac{dx^{\mu}}{ds} \frac{dx_{\mu}}{ds}} + q A_{\mu}(x(s)) \frac{dx^{\mu}}{ds} \right\} ds \right] \quad (2.41)$$

The Wiener measure $[dx^{\mu}(s)]$ is a measure, or “volume form” on the space of all maps from $\mathcal{D}$ into $\mathbb{M}^{1,d-1}$. Defining this is nontrivial, but all we need to know is that it exists and is well-defined. Let us verify that this is a reasonable action principle:

The stationary variation of the classical action: ⁴

$$S = \int_{\mathcal{D}} \left\{ mc \sqrt{-\frac{dx^{\mu}}{ds} \frac{dx_{\mu}}{ds}} + q A_{\mu}(x(s)) \frac{dx^{\mu}}{ds} \right\} ds \quad (2.42)$$

under $x^{\mu}(s) \rightarrow x^{\mu}(s) + \delta x^{\mu}(s)$ leads to the equation of motion:

$$mc \frac{d}{ds} \left(\frac{1}{\Gamma} \frac{d}{ds} x_{\mu} \right) = -q F_{\mu\nu}(x(s)) \frac{dx^{\nu}}{ds} \quad (2.43)$$

where $\Gamma := \sqrt{-\frac{dx^{\mu}}{ds} \frac{dx_{\mu}}{ds}}$. In order to relate this to the familiar equations of motion we define the proper time τ by

$$cd\tau = \Gamma ds \quad (2.44)$$

In particular, taking $x^0 = ct$ and $s = t$ this gives $\Gamma = c\gamma = c\sqrt{1 - v^i v^i / c^2}$ and

$$d\tau = \gamma dt. \quad (2.45)$$

Then:

$$m \frac{d^2}{d\tau^2} x_{\mu} = -\frac{q}{c} F_{\mu\nu}(x(\tau)) \frac{dx^{\nu}}{d\tau} \quad (2.46)$$

Now

$$F = E \wedge dx^0 + B \quad (2.47)$$

so using special properties of $d = 3 + 1$ dimensions:

$$F_{i0} = E_i \quad F_{ij} = \epsilon_{ijk} B_k \quad (2.48)$$

⁴For details on the computation, if you get stuck you could consult, for examples, J.D. Jackson, *Classical Electrodynamics*, 2nd. ed., section 12.1 or L.D. Landau and E.M. Lifshitz, *The Classical Theory of Fields*, section 16.

and hence the $\mu = 0$ component of (2.43) becomes

$$\frac{d}{dt} \left(\frac{mc^2}{\gamma} \right) = -cq v^i E_i = -cq \vec{v} \cdot \vec{E} \quad (2.49)$$

where $\vec{v} = \frac{d\vec{x}}{dt}$ and the $\mu = i$ component becomes:

$$\frac{d}{dt} \left(\frac{m\vec{v}}{\gamma} \right) = -qc\vec{E} - q\vec{v} \times \vec{B} \quad (2.50)$$

These are the standard equations of motion for a charged particle in an electromagnetic field. (Recall that to return to standard cgs conventions we should rescale charges and fields by a factor of $c^{-1/2}$.)

Thus, (2.42) is a good action principle. Nevertheless, the expression (2.38) has a very peculiar property: It is not *manifestly* gauge invariant. It is, in fact, gauge invariant, when suitably interpreted:

If we make a gauge transformation $A_\mu \rightarrow \tilde{A}_\mu = A_\mu + \partial_\mu \chi$ then it changes by

$$\begin{aligned} \exp[i\frac{q}{\hbar} \int_\gamma \tilde{A}] &= \exp[i\frac{q}{\hbar} \int_\gamma A] \exp[i\frac{q}{\hbar} \int_\gamma d\chi] \\ &= \exp[i\frac{q}{\hbar} \int_\gamma A] \exp[i\frac{q}{\hbar} \int_{\mathcal{D}} \partial_\mu \chi(x(s)) \frac{dx^\mu}{ds}] \\ &= \exp[i\frac{q}{\hbar} \int_\gamma A] \exp[i\frac{q}{\hbar} \chi(x(s_f))] \exp[-i\frac{q}{\hbar} \chi(x(s_i))] \end{aligned} \quad (2.51)$$

Now:

1. It is not quite true that $\chi(x)$ must be a single-valued function on spacetime. If the gauge group is $U(1)$ (and not its universal cover) then $e^{i\frac{q}{\hbar}\chi(x)}$ must be single valued. If the gauge group is the additive group $\mathbb{R}$ (the universal cover of $U(1)$) then we must impose the more stringent condition that $\chi(x)$ be single-valued. In any case, if $\mathcal{D} = S^1$ then for a single-valued wavefunction $x_i = x_f$ and the factors cancel.
2. If $\mathcal{D}$ is an interval, half-interval, or the real line then the path integral is interpreted as a kernel $K(x(s_i), x(s_f))$ so that given a wavefunction $\psi_i(x)$

$$\psi_f(x_f) = \int d\vec{x}_i \psi_i(x_i) K(x_i, x_f) \quad (2.52)$$

where $x_i = x(s_i)$ and $x_f = x(s_f)$ are the boundary values in the path integral. Then under a gauge transformation the wavefunction of a charged particle of charge q transforms according to

$$\psi(x) \rightarrow e^{i\frac{q}{\hbar}\chi(x)} \psi(x) \quad (2.53)$$

We conclude that the action (2.42) is indeed suitably gauge-invariant.

Remark: The action (2.42) has a nice generalization to $x : \mathcal{D} \rightarrow \mathcal{S}$ where $\mathcal{S}$ is a (pseudo-) Riemannian manifold. If, in local coordinates the metric on $\mathcal{S}$ is written as $G_{\mu\nu}(x)dx^\mu \otimes dx^\nu$ then we use

$$S = \int_{\mathcal{D}} \left\{ mc \sqrt{-\frac{dx^\mu}{ds} \frac{dx^\nu}{ds} G_{\mu\nu}(x(s))} + q A_\mu(x(s)) \frac{dx^\mu}{ds} \right\} ds \quad (2.54)$$

in the pseudo-Riemannian case and

$$S = \int_{\mathcal{D}} \left\{ mc \sqrt{\frac{dx^\mu}{ds} \frac{dx^\nu}{ds} G_{\mu\nu}(x(s))} + q A_\mu(x(s)) \frac{dx^\mu}{ds} \right\} ds \quad (2.55)$$

in the Riemannian case. In both cases, the kinetic term is the *induced line element* on the worldline of the particle.

Exercise

Derive the equation of motion following from (2.54) and (2.55).

Exercise *Point particle action as 0 + 1-dimensional “quantum gravity”*

a.) Show that the action (2.34) for a charged particle moving in an electromagnetic field has a gauge invariance under reparametrizations $s \rightarrow f(s)$ where $f(s)$ is a monotonically increasing differentiable function of s .

b.) Introduce a metric on the domain $\mathcal{D}$, say it is $g_{ss}(s)(ds)^2$. Since $g_{ss} > 0$ we can define a positive squareroot so the length element is $e(s)ds$. (This would be called an *einbein* in general relativity.) Suppose $x : \mathcal{D} \rightarrow \mathcal{S}$ is a map into a pseudo-Riemannian manifold with metric $G_{\mu\nu}(x)$ of signature $(-1, +1^{d-1})$. Consider the action:

$$S = \frac{1}{2} m \int_{\mathcal{D}} \left(-g^{ss}(s) \frac{dx^\mu}{ds} \frac{dx^\nu}{ds} G_{\mu\nu}(x(s)) + c^2 \right) e(s) ds \quad (2.56)$$

$G_{\mu\nu}$ has signature mostly plus. Show that the einbein can be eliminated by algebraic equations of motion to produce the action for a particle moving in a spacetime $\mathcal{S}$.

c.) Verify that the action is invariant under diffeomorphisms $s \rightarrow f(s)$ provided $e(s)ds$ transforms like a line element.

d.) Show that with a suitable rescaling of e one can take an $m \rightarrow 0$ limit preserving a good kinetic term.

Interpretation: We are doing quantum gravity in 0 + 1 dimensions with a “graviton” $e(s)ds$ and “matter fields” $x^\mu(s)$ with a possibly nonlinear target. The term $\int m e ds$ is like the cosmological constant. The coupling $q \int A_\mu dx^\mu$ is independent of the metric and hence is a “topological term.”

For a computation of the propagator of a particle through spacetime from this point of view see, for example, ⁵

⁵A. Cohen, G. Moore, P. Nelson, and J. Polchinski, “An off-shell propagator for string theory,” Nucl. Phys. **B267**(1986)143

2.6 A charged particle in a flat gauge field

When $\mathcal{D} = S^1$ there is another, very beautiful way to see that (2.38) is gauge invariant. The argument works in some important special cases. Suppose the image $\gamma \subset \mathcal{S}$ of $\mathcal{D} = S^1$ is a closed curve in spacetime and we can use Stokes' theorem (integration by parts) to rewrite:

$$\exp[i\frac{q}{\hbar} \int_{\gamma} A] = \exp[i\frac{q}{\hbar} \int_D F] \quad (2.57)$$

where $D \subset \mathcal{S}$ is a disk whose boundary is γ .

However, there can be situations in which γ does not bound a disk. There can even be situations where γ does not bound a disk and $F = 0$ in $\mathcal{S}$ and yet (2.38) is not $= 1$!

As an example, consider a infinitely thin solenoid carrying a flux Φ . Let us say that it runs along the x^3 -axis in $\mathbb{R}^3$. Suppose it is impenetrable to our particle. Then we must take $\mathcal{S} = \mathbb{R} \times (\mathbb{R}^3 - \mathcal{Z})$ where $\mathcal{Z} = \{(0, 0, x^3) : x^3 \in \mathbb{R}\}$ is the z -axis. Note that $F = 0$ in $\mathcal{S}$, but $A = \frac{\Phi}{2\pi} d\phi$ is nonzero, and for general Φ , we cannot gauge A to zero by a single-valued gauge transformation! Thus, *even though $F_{\mu\nu} = 0$ the gauge field is nontrivial and cannot be gauged to zero.* Such a gauge field is called a (nontrivial) *flat gauge field*.

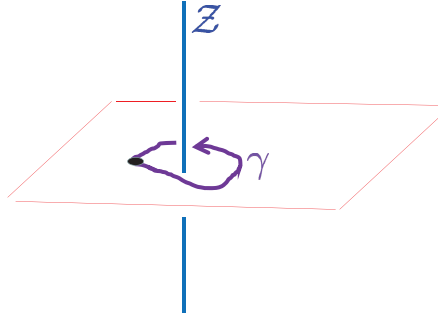


Figure 5: A solenoid carrying flux Φ runs along the z -axis $\mathcal{Z}$ and is cloaked by an impenetrable barrier. A particle moves around the solenoid on a path γ . The time direction is suppressed in this figure.

Now imagine that the worldline of the charged particle makes a closed loop around the solenoid as in Figure 5. We can compare the original wavefunction to the final one. For simplicity suppose that $A_0 = 0$ so that we can just consider the projection of the worldline into space. The factor (2.38) becomes

$$\exp[i\frac{q\Phi}{\hbar}] \quad (2.58)$$

In fact, if the particle moves arbitrarily slowly around the loop then one can show (using the quantum adiabatic theorem) that this is the only phase factor the particle picks up. But this factor need not be unity! As opposed to the gauge freedom (2.53) this phase

change of the wavefunction is physical and would have an effect, for example, in changing phase interference in the famous double-slit experiment.⁶

Again, we see an interesting version of topological invariance in physics: The precise choice of loop γ does not matter for the phase (2.58) in the sense that it is unchanged under continuous deformations of γ which do not cross the solenoid.

Exercise

Suppose γ loops around the solenoid n times. What is the phase shift?

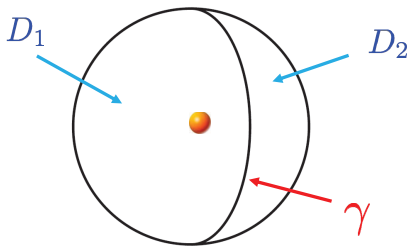


Figure 6: A closed path $\gamma \subset \mathbb{R}^3 - \{0\}$ can be bounded by two different disks D_1, D_2 such that $D_1 \cup D_2$ is a closed 2-cycle linking the origin.

2.7 Dirac's argument for quantization

The quantization of the angular momentum (2.35) is *not* how Dirac originally discovered his quantization law in his spectacular paper of 1931.⁷ Rather, he studied the quantum mechanics of a charged particle of charge moving in the field of a hypothetical magnetic monopole. Because of the singularity of the field at the origin that means we consider the charged particle as moving in $\mathbb{R}^3 - \{0\}$.

If the path γ is a closed path it is not obvious that (2.38) is well-defined and gauge invariant. However, we can use Stokes' theorem to argue that it is gauge invariant: Choose a disk D whose boundary is the path γ . Then we can say that

$$\exp[i\frac{q}{\hbar} \int_{\gamma} A] = \exp[i\frac{q}{\hbar} \int_D dA] = \exp[i\frac{q}{\hbar} \int_D F] \quad (2.59)$$

⁶For an excellent account of this see S. Coleman, "The magnetic monopole: 50 years later."

⁷In this paper he predicted the existence of *three* new particles, the anti-electron, the anti-proton, and the magnetic monopole. At the time, predicting new particles was *not* something that a theorist was expected to do. The reference is P.A.M. Dirac, "Quantised Singularities in the Electromagnetic Field," Proc. Roy. Soc. A 133, 60.

On the other hand, as shown in Figure 6, the choice of the disk D is ambiguous. In order for the phase of the electron wavefunction to be well-defined we must have

$$\exp[i\frac{q}{\hbar} \int_{D_1} F] = \exp[i\frac{q}{\hbar} \int_{D_2} F] \quad \Rightarrow \quad \exp[i\frac{q}{\hbar} \int_{\Sigma} F] = 1 \quad (2.60)$$

where $\Sigma = -D_1 \cup D_2$ is the closed two-surface enclosing the monopole and the minus sign refers to orientation. But the integral $\int_{\Sigma} F$ just measures the total magnetic flux through Σ and can be evaluated by Gauss' law.

More precisely, the field of a static magnetic charge g is

$$F = \frac{g}{2} \epsilon_{ijk} \frac{x^i dx^j dx^k}{r^3} \quad (2.61)$$

and the integral on a surface enclosing the origin is just $4\pi g$. Thus, the phase of the wavefunction is well-defined if

$$\frac{4\pi g q}{\hbar} \quad (2.62)$$

is an integral multiple of 2π , so that

$$gq = \frac{1}{2} n \hbar \quad n \in \mathbb{Z} \quad (2.63)$$

The generalization to an arbitrary pair of dyons can be shown to be

$$g_1 q_2 - g_2 q_1 = \frac{1}{2} n \hbar \quad n \in \mathbb{Z} \quad (2.64)$$

The topology implied here is rather more subtle: It has to do with the topology of fiber bundles and how that is measured by connections on those bundles.

Exercise

Show that if D_1 can be deformed to $\tilde{D}_1$ without crossing the magnetic pole then

$$\exp[i\frac{q}{\hbar} \int_{D_1} F] = \exp[i\frac{q}{\hbar} \int_{\tilde{D}_1} F] \quad (2.65)$$

follows without using the Dirac condition.

2.8 Anyons in 2+1 dimensions

Let us now consider what happens when charged particles are constrained to live in two spatial dimensions. See

1. <http://en.wikipedia.org/wiki/2DEG>
2. S. Girvin, "The Quantum Hall Effect: Novel Excitations and Broken Symmetries," arXiv:cond-mat/9907002

for a description of some experimental realizations approximating this idealized situation.

Now, it is interesting if our plane also has thin solenoids of flux Φ piercing it. We can imagine a situation in which the flux cannot spread out, so they behave like particles in $2+1$ dimensions as well. A good example of how this can happen is in a superconductor. A nice way to understand superconductivity is that it is a theory of electromagnetism where the $U(1)$ gauge theory symmetry is spontaneously broken by the vacuum expectation value of a charge two field representing the Cooper pairs. The flux tubes are regions of normal phase, where the photon is massless. The superconductor is a region where the photon gets a mass. The flux cannot spread out.

Now imagine that - for some unspecified reason - a particle of charge q binds to such a solenoid-particle. We label the boundstate by (q, Φ) . These $2+1$ dimensional analogs of dyons have some very curious properties.

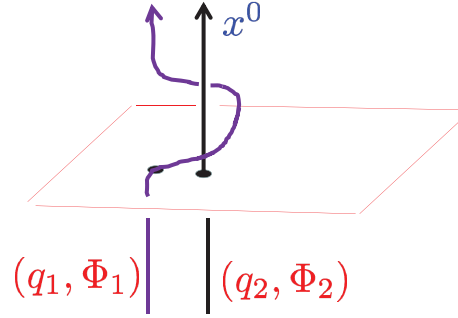


Figure 7: A boundstate (q_1, Φ_1) moves very slowly counterclockwise around a boundstate (q_2, Φ_2) . Only the topology of the path matters in computing the change of phase of the wavefunction. Do not confuse the vertical direction with the z -axis. The vertical direction now represents the time direction.

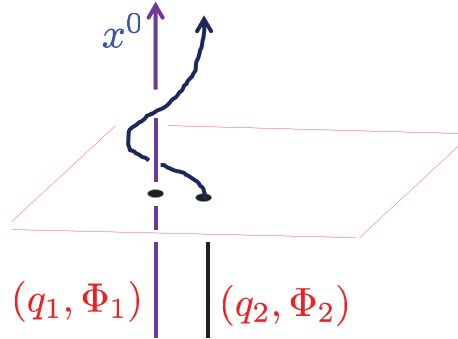


Figure 8: A topologically equivalent formulation of the path in Figure 7. This makes it clear that the boundstate (q_2, Φ_2) also moves counterclockwise around (q_1, Φ_1) .

Let us move a particle (q_1, Φ_1) very slowly around a particle (q_2, Φ_2) as in Figure 7. Applying the formula (2.38) the wavefunction picks up a phase $\exp[\frac{i}{\hbar c}(q_2\Phi_1)]$. Note that

this does not depend on the exact shape of the trajectories, only that one particle circles around the other. At the same time, there is a phase change because particle (q_2, Φ_1) loops around the flux Φ_1 . Indeed we could deform Figure 7 to Figure 8. Altogether then, the wavefunction of the pair of particles changes by

$$\exp\left[\frac{i}{\hbar}(q_2\Phi_1 + q_2\Phi_2)\right] \quad (2.66)$$

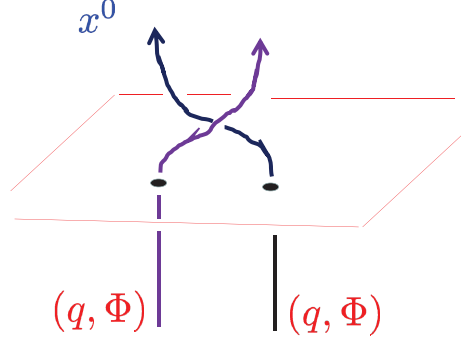


Figure 9: A pair of identical particles (q, Φ) are exchanged.

Now let us consider a pair of identical particles which are exchanged as in Figure 9. The net phase change is just $\exp[i\frac{q\Phi}{\hbar}]$. But since we have exchanged identical particles we can interpret this as a statistics phase. Unlike the case of particles in $3 + 1$ dimensions, in the present case the statistics phase can be *any* phase. Such particles are called *anyons*.⁸

It is interesting to check the relation between spin and statistics.

We now apply the general formula (2.7) to the angular momentum in $d = 2 + 1$ dimensions. Here there is just the one generator $J = J_{12}$.

In $2+1$ dimensions the solenoid contributes $F_{ij} = \epsilon_{ij}\Phi\delta^2(x)$, (here i, j run from 1 to 2 and $\epsilon_{12} = +1$). From (2.12) the electric particle at $\vec{R}$ contributes an electric field

$$\vec{E} = 2q \frac{\vec{x} - \vec{R}}{|\vec{x} - \vec{R}|^2} \quad (2.67)$$

so $F_{0j} = -2q(x - R)^j/|\vec{x} - \vec{R}|^2$. Therefore the momentum density is

$$T_{0i} = F_{0k}F_{ik} = \frac{q\Phi}{2\pi}\epsilon_{ij}\frac{R_j}{R^2}\delta^{(2)}(\vec{x}) \quad (2.68)$$

Thanks to the δ -function in the integral is easily done and we find

$$J_{12} = \frac{q\Phi}{2\pi} \frac{\vec{a} \cdot \vec{R}}{\vec{R} \cdot \vec{R}} \quad (2.69)$$

It is now amusing to check the relation of spin and statistics:

⁸The possible existence of anyons was pointed out by Leinaas and Myrheim in 1977. The term “anyon” was invented in F. Wilczek, “Quantum Mechanics of Fractional-Spin Particles”. Physical Review Letters 49 (14): 957959.

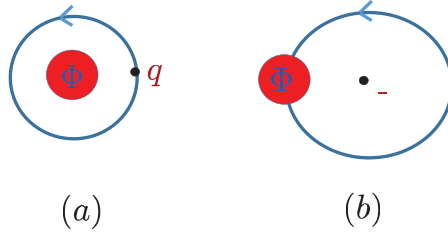


Figure 10: In (a) the charge q is slowly moved around the fluxon Φ and the wavefunction acquires an Aharonov-Bohm phase. In (b) we perform a rotation by 2π centered on q and the wavefunction of the electromagnetic field acquires a phase. These two phases are the same.

1. If we slowly rotate the particle around the flux in a counterclockwise fashion then the wavefunction picks up a phase $\exp[iq\Phi/\hbar]$.
2. On the other hand, if we rotate the flux around the particle then the wavefunction should change by $\exp[2\pi i J/\hbar]$. Taking $\vec{a} = \vec{R}$ in (2.69) we get the same phase:

$$\exp[2\pi i J/\hbar] = \exp[iq\Phi/\hbar] \quad (2.70)$$

Remark *Spin-statistics theorem:* The important property used in proving the spin-statistics theorem is the existence of an analytic continuation to Euclidean space.

Here are some sources for more material about anyons:

1. There are some nice lecture notes by John Preskill, which discuss the potential relation to quantum computation and quantum information theory: <http://www.theory.caltech.edu/~preskill/p>
2. For a reasonably up-to-date review see A. Stern, "Anyons and the quantum Hall effect A pedagogical review". *Annals of Physics* 323: 204; arXiv:0711.4697v1.
3. A. Lerda, *Anyons: Quantum mechanics of particles with fractional statistics* Lect. Notes Phys. M14 (1992) 1-138
4. Other refs include: A. Khare, *Fractional Statistics and Quantum Theory*, and G. Dunne, *Self-Dual Chern-Simons Theories*.

2.9 The moral of the examples

Simple physical computations can lead to topologically invariant quantities.

There are many other and more sophisticated examples:

1. Path integrals in topological field theories compute interesting topological invariants of highly nontrivial manifolds such as manifolds consisting of "moduli spaces" - parameter spaces of solutions to important differential equations in physics.

2. Some simple expressions in condensed matter physics have interesting topological interpretations. As a famous example of this, the Kubo formula for the Hall conductance of a noninteracting gas of electrons in a magnetic field, confined to a plane, has the interpretation of the first Chern class of a line bundle over a Brillouin torus. This is the starting point for the relation of the quantum Hall effect to topology.

3. Some basic definitions in topology

Topology is the study of *invariance of a quantity under continuous deformations*. This is often (but not always) the origin of topological quantities in physics. We will now formalize this notion mathematically.

We assume you have seen basic point-set topology. For more detail, see the refs at the end. The following definitions are essential: ⁹

3.1 Definition of a topology

Definition Let X be any set. A *topology on X* is a collection $\mathfrak{T}$ of subsets of X (called the set of *open sets of the topology* $\mathfrak{T}$) which satisfies the following conditions:

1. $\emptyset$ and X are members of $\mathfrak{T}$
2. The union of the members of an arbitrary subset of $\mathfrak{T}$ is a member of $\mathfrak{T}$.
3. The intersection of a finite number of members of $\mathfrak{T}$ is a member of $\mathfrak{T}$.

Some frequently used terminology:

1. If $x \in X$ then an open set $U \in \mathfrak{T}$ containing x is also called a *neighborhood of x* .
2. If $C \subset X$ is a subset of X so that $X - C$ is open then C is said to be *closed*. Note that both $\emptyset$ and X are both open and closed.
3. If $A \subset X$ is any subset the *closure of A* , denoted $\bar{A}$, is the intersection of all closed sets containing A . It is the smallest closed set containing A .

Two extreme examples of topologies on X are the following: We could take $\mathfrak{T}$ to be the set of all subsets of X . This is called the *discrete topology*. On the other hand, we could take $\mathfrak{T}$ to consist of a set with two members, namely $\emptyset$ and X . That defines the *trivial topology*. In general, neither of these topologies is useful in physics, and physicists do not generally worry about the precise definition of the topology of the spaces they are working with, since it is often intuitively obvious what is topology $\mathfrak{T}$ is the “correct” one. For example, when working with field theory in d -dimensional Minkowski space $\mathbb{M}^{1,d-1}$ there is an obvious topology inherited from $\mathbb{R}^d$. One place where this is not the case is in

⁹A primary source for the review material on point-set topology has been J.R. Munkres, *Topology, a first course*, Prentice Hall, 1975

the topology of operator algebras, as we will see in Chapter 3. Therefore, it is useful to have the following definition:

Definition If $\mathfrak{T}$ and $\mathfrak{T}'$ are two topologies on a set X then we say that $\mathfrak{T}'$ is *finer*, or *stronger*, than $\mathfrak{T}$ if $\mathfrak{T}'$ contains $\mathfrak{T}$ as a subset. Equivalently we say that $\mathfrak{T}$ is *coarser*, or *weaker* than $\mathfrak{T}'$.

Thus, the finest, or strongest topology on X is the discrete topology. The coarsest, or weakest topology on X is the trivial topology.

One useful way to define the topology on a set X is to give a *subbasis* $\mathfrak{B}$ of open sets. This is simply a collection of subsets of X whose union equals X . The *topology generated by* $\mathfrak{B}$ is the set of all unions of finite intersections of elements of $\mathfrak{B}$. Put differently, the topology generated by $\mathfrak{B}$ is the coarsest, or weakest topology for which all the elements of $\mathfrak{B}$ are open sets.¹⁰

One natural way in which a subbasis is useful is in defining the metric topology on a metric space.

Definition A *metric* on a set X is a map $d : X \times X \rightarrow \mathbb{R}$ such that

1. $d(x_1, x_2) = d(x_2, x_1)$
2. $\forall x_1, x_2, x_3 \in X, d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3)$
3. $d(x_1, x_2) \geq 0$ for all $x_1, x_2 \in X$
4. $d(x_1, x_2) = 0$ iff $x_1 = x_2$

A set X equipped with a metric is called a *metric space*.

Warning: The “Minkowski metric” is not a metric in the above sense!

Definition The *metric topology* on a metric space (X, d) is the topology generated by the subbasis of open balls:

$$B(x_0, \epsilon) := \{x \in X \mid d(x, x_0) < \epsilon\} \quad (3.1)$$

Exercise

If $(X, \mathfrak{T}_X)$ is a topological space and $Y \subset X$ is any subset, so that $\{U \cap Y \mid U \in \mathfrak{T}_X\}$ defines a topology on Y . It is called the *subspace topology* or *induced topology*.

¹⁰One naturally asks what a *basis* for a topology would be. This is a collection $\mathfrak{B} \subset \mathfrak{T}$ of open sets such that every open set is a union of elements of $\mathfrak{B}$.

Exercise

Formulate a definition of a topology on a space X in terms of the set of *closed sets*.

3.2 Continuity and convergence

Now, finally, we can define a *continuous function* between two topological spaces:

Definition A function $f : X \rightarrow Y$ between two sets with topologies $\mathfrak{T}_X$ and $\mathfrak{T}_Y$ is said to be *continuous* if, for all open sets in Y , that is, for all $V \in \mathfrak{T}_Y$, the inverse image $f^{-1}(V)$ is an open set in X , that is, $f^{-1}(V) \in \mathfrak{T}_X$.

A closely related notion is the *convergence of a sequence* $\{x_n\}_{n \geq 1}$ of points in X with a topology $\mathfrak{T}_X$. We say that such a sequence converges to $x \in X$ if for every neighborhood $U \in \mathfrak{T}_X$ of x there exists an N so that for all $n \geq N$ we have $x_n \in U$.

Remarks

1. For *any* subset $W \subset Y$ the set of points in X which do not map to W under f is $f^{-1}(Y - W) = X - f^{-1}(W)$. It follows that an equivalent definition is that f is continuous iff the inverse image of closed sets is closed.
2. A common mistake is to assume that if f is continuous and U is open then $f(U)$ is open. (Continuous functions which do satisfy this property are called *open*.) A trivial counterexample is to take the real-valued function which maps all of $\mathbb{R}^n$ to a point. It is also a mistake to assume that a continuous function will map a closed set into a closed set. A counterexample is $f : \mathbb{R} \rightarrow U(1) \times U(1)$ defined by $f(x) = (e^{2\pi i x}, e^{2\pi i \alpha x})$ where α is an irrational real number.
3. We can relate the notion of continuity and convergence of a sequence as follows: Let $\widehat{\mathbb{Z}}_+ = \mathbb{Z}_+ \cup \{\infty\}$ have the topology determined by the subbasis of sets $B_N := \{n | n \geq N\}$. Then a sequence x_n in X converges to x iff the function $f : \widehat{\mathbb{Z}}_+ \rightarrow X$ defined by $f(n) = x_n$ and $f(\infty) = x$ is continuous.
4. There are a number of “separation axioms” one can put on topologies. One important one is the *Hausdorff condition*. This is the condition that one can separate points by open sets. That is: For all pairs $x_1, x_2 \in X$ of distinct points there exist open sets U_1, U_2 with $x_1 \in U_1$, $x_2 \in U_2$, but $U_1 \cap U_2 = \emptyset$. The condition comes up in physics sometimes because in forming quotient spaces, for example when dividing by some gauge symmetry, one sometimes encounters a difficulty that the quotient space with the quotient topology is not Hausdorff. This usually means one of two things: Some important degree of freedom was overlooked or one should be working with “noncommutative geometry.” For some examples of non-Hausdorff topological spaces consider:

Example A: The trivial topology on any set X with more than one element.

Example B: The set of equivalence classes of points on the unit circle with an equivalence relation determined by choosing an irrational real number α . The equivalence relation is $z_1 \sim z_2$ if $z_1 = e^{2\pi i n \alpha} z_2$ where $n \in \mathbb{Z}$. The set of equivalence classes inherits a topology (see Section §5 below) and this topology is easily seen to be non-Hausdorff. We will see other examples when we discuss group quotients in Chapter ****

5. In Chapter 1 we discussed at length the notion of groups. We can now combine the notion of a topological space with that of a group to define a *topological group*. This is a group G which is also endowed with a topology. But a group with a topology is not necessarily a topological group! We must make sure that the two mathematical structures are compatible and interact well together. In particular for G to be a topological group we demand that the multiplication map $G \times G \rightarrow G$ and the inversion map $G \rightarrow G$ which takes $g \rightarrow g^{-1}$ are both continuous maps. An important class of topological groups are the *Lie groups* studied in **** below. However, not all topological groups are Lie groups. A simple example would be the rational numbers.

Exercise

Suppose that $f : X \rightarrow Y$ is a continuous map between topological spaces $(X, \mathfrak{T}_X)$ and $(Y, \mathfrak{T}_Y)$. Show that

- a.) If $\mathfrak{T}'_Y$ is weaker than $\mathfrak{T}_Y$ then f is continuous as a function to $(Y, \mathfrak{T}'_Y)$.
 - b.) If $\mathfrak{T}'_X$ is stronger than $\mathfrak{T}_X$ then f is continuous as a function from $(X, \mathfrak{T}'_X)$.
 - c.) If $\mathfrak{T}'_X$ is weaker than $\mathfrak{T}_X$ then a sequence $\{x_n\}$ in X which converges to x in $\mathfrak{T}_X$ also converges to x in $\mathfrak{T}'_X$.
-

Exercise

Suppose the topology of Y is generated by a subbasis $\mathfrak{B}$. Show that $f : X \rightarrow Y$ is continuous iff $f^{-1}(S) \subset X$ is open for every element $S \in \mathfrak{B}$.¹¹

3.3 Homeomorphism

Definition : A *homeomorphism* $\phi : X \rightarrow Y$ of two topological spaces X, Y is a 1-1 continuous map with continuous inverse.

Remarks:

1. In the definition of homeomorphism it is necessary to say that the bijective continuous map has an inverse. For an example of a bijective continuous map whose inverse is not

¹¹Hint: If $V = S_1 \cap \cdots \cap S_n$ then $f^{-1}(V) = f^{-1}(S_1) \cap \cdots \cap f^{-1}(S_n)$.

continuous, take any set X and choose two topologies $\mathfrak{T}_X$ and $\mathfrak{T}'_X$. Let $f : X \rightarrow X$ be the identity map. This is certainly bijective! If we regard it as a map of topological spaces $(X, \mathfrak{T}_X) \rightarrow (X, \mathfrak{T}'_X)$ then if $\mathfrak{T}_X$ is stronger than $\mathfrak{T}'_X$ the map will be continuous, but if $\mathfrak{T}_X$ is strictly stronger than $\mathfrak{T}'_X$ then the inverse map will not be continuous.

2. A continuous map $f : X \rightarrow Y$ of topological spaces is a *topological embedding* if f is a homeomorphism of X with its image $f(X) \subset Y$ where $f(X)$ carries the subspace topology inherited from Y .

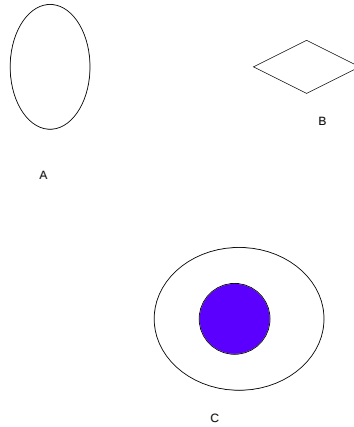


Figure 11: The regions inside the curves A and B are homeomorphic to each other, but not to the annular region described by C.



Figure 12: A complicated projection of a knot in $\mathbb{R}^3$. Is it equivalent to the unknot?

Examples

1. Consider the square, disk, and annulus. The square and disk are homeomorphic. They are not homeomorphic to the annulus.
2. $\mathbb{R}^n$ is not homeomorphic to $\mathbb{R}^m$ for $n \neq m$.

3. Define the n -dimensional sphere for $n \geq 0$ to be:

$$S^n \equiv \{\vec{x} \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} (x^i)^2 = 1\} \quad (3.2)$$

It is not homeomorphic to $\mathbb{R}^n$.

The above three statements are intuitively obvious. But it is perhaps not immediately obvious how to prove them. One way to prove such statements is by finding a topological invariant.

Definition A *topological invariant* is any quantity assigned to topological spaces which depends only on the homeomorphism equivalence class of the space.

Example: Consider a knot in $C \subset \mathbb{R}^3$. C is homeomorphic to S^1 , but its embedding in $\mathbb{R}^3$ can be complicated. So, for different C , $\mathbb{R}^3 - C$ will not be homeomorphic. It can be hard to recognize if the knot is equivalent to the “unknot” in $\mathbb{R}^3$. See for example Figure 12. Similarly, if we have several embedded copies of S^1 in $\mathbb{R}^3$ then we have a link. If we move closed smooth curves C_1, C_2 in $\mathbb{R}^3$ around so that they do not cross then the spaces $\mathbb{R}^3 - [C_1 \cup C_2]$ are homeomorphic. If they do cross then the resulting spaces might or might not be homeomorphic. The linking number $L(C_1, C_2)$ is one topological invariant. While many interesting invariants are known, finding a complete set of topological invariants for a knot or link in $\mathbb{R}^3$ is an unsolved problem. It is conjectured that the collection of Vassiliev “finite-type invariants” or the collection of perturbative Chern-Simons invariants forms a complete set of invariants for knots. These are fancier integral formulae generalizing the Gauss linking formula.

Two simple homeomorphism invariants are *compactness* and *connectedness*. We discuss those next.

Exercise

Prove that $X \sim Y$ if X, Y are homeomorphic is an equivalence relation.

3.4 Compactness

Definition :

- a.) A space X is *compact* if every covering by open sets has a finite sub-covering.
- b.) A space X is *locally compact* if at every point $x \in X$ there is a compact set K containing a neighborhood of x .

Examples:

1. A continuous function maps compact sets to compact sets. Thus, compactness is a homeomorphism invariant.
2. **Warning:** It is not true in general that the inverse image under a continuous function of a compact set is compact. Continuous functions that have the property that $f^{-1}(K)$ for all compact subsets K of the codomain are said to be *proper*.
3. One can show that a subset of $\mathbb{R}^n$ is compact iff it is closed and bounded. For example, the closed interval $[0, 1]$ is compact and the open interval $(0, 1)$ is not. These are not homeomorphic.
4. S^n is compact, $\mathbb{R}^n$ is not. Therefore, they are not homeomorphic. Stereographic projection shows that $S^n - \{point\}$ is homeomorphic to $\mathbb{R}^n$.
5. The closed disk $\{z ||z| \leq R\}$ is compact and the punctured disk where we cut out $z = 0$ is not. These are not homeomorphic.
6. Almost all spaces one normally works with in physics are locally compact. An example of a space which is not locally compact is the set of rational numbers $\mathbb{Q}$ with the topology induced from that of the real line.

Remark: A frequent source of technical trouble and subtlety in field theory is non-compactness of spaces of functions, or spaces of solutions to differential equations.

3.5 Connectedness

Definition A topological space X is said to be *disconnected* if there exist two nonempty disjoint open sets U, V so that $U \cup V = X$. If X is not disconnected it is said to be *connected*.

If $f : X \rightarrow Y$ is a continuous map then if X is connected then so is $f(X)$. However, the inverse image of a connected set need not be connected! Just think of some many-to-one maps. Nevertheless, if f is a homeomorphism then since f^{-1} is connected, X and Y are either both connected or disconnected. Thus, connectedness is a homeomorphism invariant.

Exercise

Show that X is connected iff the only elements of $\mathfrak{T}_X$ that are both open and closed are $\emptyset$ and X .¹²

We can define an equivalence relation on X by saying that $x_1 \sim x_2$ if there is a connected set $U \subset X$ which contains both x_1 and x_2 . The equivalence classes under this

¹²Ans: Suppose V is both open and closed and is not X or the empty set. Then V is nonempty, by assumption and $U = X - V$ is open (since V is closed) and nonempty, since $V \neq X$. But $U \cup V = X$.

relation are called the *connected components* of X . The *set of connected components* is usually denoted $\pi_0(X)$. It is a basic homeomorphism invariant of X .

Example: *Spheres.* The 0-dimensional sphere is disconnected: $\pi_0(S^0) = \{p_+, p_-\}$ has two elements but $\pi_0(S^n)$ has one element for $n > 0$. This is often the source of special phenomena in low dimensions.

Remark: There are many refinements of the notion of connectedness: A *path in X* is a continuous map $[0, 1] \rightarrow X$ and a space is *path connected* if any two points can be connected by a path. Topology books show that a path connected space is connected, but that the converse is false. There are also notions of “arcwise connected,” “hyperconnected,” and “locally connected.” We will not need to worry about such refinements.

Exercise

Show that if X has more than one element and the discrete topology it is disconnected.

4. Topologies on spaces of functions

In field theory we are often concerned with *function spaces* or *spaces of maps* between two spaces X and Y . Typically, X is some spacetime and Y is some kind of “target space.” We need a notion of when two field configurations are “close.”

The most primitive thing we can do is consider the space of *all* maps $\text{Map}(X, Y)$. This space can be given a topology, known as the *point-open topology* or *topology of pointwise convergence*. This is defined by giving a subbasis, labeled by a pair consisting of a point $x \in X$ and an open set of Y , i.e. $U \in \mathfrak{T}_Y$. Then we define:

$$B^{p.o.}(x, U) := \{f \mid f(x) \in U\} \quad (4.1)$$

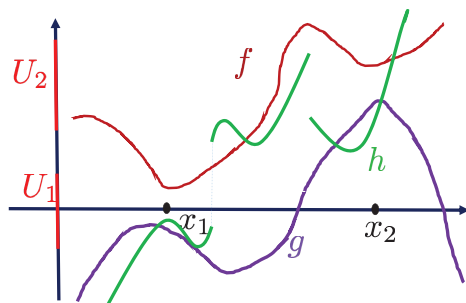


Figure 13: The three functions f, g, h are in the same open neighborhood $B^{p.o.}(x_1, U_1) \cap B^{p.o.}(x_2, U_2)$ in the point-open topology on $\text{Map}(\mathbb{R}, \mathbb{R})$.

This is a rather coarse or weak topology. The open sets are made by taking finite intersections

$$B^{p.o.}(x_1, U_1) \cap \cdots \cap B^{p.o.}(x_n, U_n) \quad (4.2)$$

Thus, two functions f, g are “close” in the sense that they are in the same open set if they both map x_j into U_j for $j = 1, \dots, n$. They could look very different in an intuitive sense as shown in Figure 13.

In the point-open topology one can show ¹³ that a sequence of functions $f_n : X \rightarrow Y$ converges to $f : X \rightarrow Y$ iff for each $x \in X$, the image $f_n(x)$ converges to $f(x)$. A clear disadvantage of this topology now becomes apparent: A sequence of *continuous* functions need not converge to a continuous function. For example consider $\text{Map}([0, 1], \mathbb{R})$ and let $f_n(x) = x^n$. For $n \rightarrow \infty$ this converges in the point-open topology to the function

$$f(x) = \begin{cases} 0 & 0 \leq x < 1 \\ 1 & x = 1 \end{cases} \quad (4.3)$$

which is not continuous.

There are two ways to try to cure this problem:

1. Put more structure on Y and use that to define a finer topology on $\text{Map}(X, Y)$.
2. Restrict to important subspaces of $\text{Map}(X, Y)$, such as the space of continuous functions $\mathcal{C}(X, Y)$ and define a different topology there.

In field theory we often do both: We restrict to spaces of functions which are suitably smooth, and we use extra structures, such as an action functional, to help define a topology on the space of fields. See examples in Section **** below.

As an example of the first procedure we put the extra structure of a metric space on Y . Thus, suppose that (Y, d) is a metric space. Then we can define the *topology of compact convergence* on $\text{Map}(X, Y)$ by choosing a subbasis of open sets to be labeled by triples of $f \in \text{Map}(X, Y)$, a compact set $K \subset X$, and $\epsilon > 0$. Then we have

$$B^{c.c.}(f, K, \epsilon) := \{g \in \text{Map}(X, Y) \mid \sup_{x \in K} d(f(x), g(x)) < \epsilon\} \quad (4.4)$$

Now we have more control: One can show ¹⁴ that in the topology of compact convergence a sequence of functions $f_n : X \rightarrow Y$ converges to $f : X \rightarrow Y$ iff for each compact set $K \subset X$ the functions f_n restricted to K converge *uniformly* to f restricted to K . Written out in terms of ϵ 's and δ 's this means: For all $\epsilon > 0$ there exists N such that for all $n > N$ and for all $x \in K$

$$d(f_n(x), f(x)) < \epsilon \quad (4.5)$$

The important point is that $f_n(x)$ and $f(x)$ are close not just at a few points, but throughout K .

¹³Munkres, Theorem 7.4.1

¹⁴Munkres, Theorem 7.4.2

As an example of the second procedure we now restrict the subspace of $\text{Map}(X, Y)$ of *continuous* functions. We denote this as $\mathcal{C}(X, Y)$. On this space we can define the *compact-open topology*. The basic open sets of a subbasis are now labeled by a compact set $K \subset X$ and an open set $U \subset Y$. They are defined by

$$B_{K,U}^{c.o.} := \{f \in \mathcal{C}(X, Y) : f(K) \subset U\} \quad (4.6)$$

Pleasantly enough, in the case when (Y, d) is a metric space $\mathcal{C}(X, Y)$ is a closed subset of $\text{Map}(X, Y)$ with the compact convergence topology, and the induced topology on $\mathcal{C}(X, Y)$ coincides with the compact-open topology (and hence does not depend on the choice of d).

¹⁵

There are two crucial facts about the compact-open topology:

The first is

Theorem If X, Y, Z are topological spaces and Y is locally compact and Hausdorff then the composition of maps defines a continuous function

$$\mathcal{C}(X, Y) \times \mathcal{C}(Y, Z) \rightarrow \mathcal{C}(X, Z) \quad (4.7)$$

Idea of proof: Let C be the composition map. The idea of the proof is to show that for every compact set $K \in \mathfrak{T}_X$ and open set $U \in \mathfrak{T}_Z$ the inverse image $C^{-1}(B^{c.o.}(K, U))$ is open. It suffices to show for every (f, g) so that $g \circ f \in B^{c.o.}(K, U)$ we can find a neighborhood $B^{c.o.}(K, V) \times B^{c.o.}(\bar{V}, U)$ of (f, g) in the inverse image $C^{-1}(B^{c.o.}(K, U))$, where V is open in Y and $\bar{V}$ is compact in Y . Then we can take the union of such open neighborhoods around all points (f, g) in the inverse image and cover $C^{-1}(B^{c.o.}(K, U))$ by a union of open sets, hence it will be open. Now, how do we find such a neighborhood V in Y ? Observe that $f(K) \subset Y$ is compact, since f is continuous. Because Y is locally compact, around any point $y \in f(K)$ there is a compact set C_y containing a neighborhood W_y of y . But then $g^{-1}(U) \cap W_y$ is also a neighborhood of y and its closure maps into U . The W_y provide a covering of $f(K)$ and hence there is a finite covering. Choosing one let V_i be the open sets. By construction $g(\bar{V}_i) \subset U$ and then we can take $V = \cup V_i$. ♠

Remarks:

1. If we take $X = pt$ then $\mathcal{C}(X, Y)$ is homeomorphic to Y and $\mathcal{C}(X, Z)$ is homeomorphic to Z . Then the map

$$Y \times \mathcal{C}(Y, Z) \rightarrow Z \quad (4.8)$$

which sends $(y, f) \mapsto f(y)$ is known as the *evaluation map*. So a corollary is that the evaluation map is continuous in the compact-open topology.

2. Do not confuse the fairly trivial statement that the composition of continuous functions is a continuous function with the more nontrivial statement that the composition map (4.7) is a continuous map of function spaces. The first, trivial, statement merely guarantees that composition maps into the subspace $\mathcal{C}(X, Z)$ of $\text{Map}(X, Z)$.

¹⁵Munkres, Theorem 7.5.1

The second crucial fact about the compact-open topology is the following: Suppose that X, Y, Z are any three sets. Given a function $F : X \times Y \rightarrow Z$ we can always define an associated function $\mathcal{F} : X \rightarrow \text{Map}(Y, Z)$ by declaring $\mathcal{F}(x)$ to be that function which sends y to $F(x, y)$. Conversely, given a function $\mathcal{F} : X \rightarrow \text{Map}(Y, Z)$ we can construct an associated function $F : X \times Y \rightarrow Z$.

If X, Y, Z are topological spaces, with Y locally compact and Hausdorff, then a function $F : X \times Y \rightarrow Z$ is continuous iff the associated function $\mathcal{F} : X \rightarrow \mathcal{C}(Y, Z)$ is continuous.

One direction is quite easy: If $\mathcal{F}$ is continuous then we recover F as the composite of $(x, y) \rightarrow (\mathcal{F}(x), y) \rightarrow \mathcal{F}(x)(y)$, where the second arrow is the evaluation map. This is a composition of continuous maps and is therefore continuous. The converse takes a little more work. See Munkres, Corollary 7.5.4.

Remarks:

1. In the point-open topology it is *not true* that $F : X \times Y \rightarrow Z$ is continuous iff $\mathcal{F} : X \rightarrow \text{Map}(Y, Z)$ is continuous. We can demonstrate this with our counterexample (4.3) above. The function $\mathcal{G} : \widehat{\mathbb{Z}} \rightarrow \text{Map}([0, 1], \mathbb{R})$ defined by taking $\mathcal{F}(n) = x^n$ for $1 \leq n < \infty$ and $\mathcal{F}(\infty)$ to be the function in (4.3) is a *continuous* function in the point-open topology. ON the other hand, the corresponding function $F : \widehat{\mathbb{Z}}[0, 1] \rightarrow \mathbb{R}$ given by

$$F(n, x) = \begin{cases} x^n & 1 \leq n < \infty \\ 0 & n = \infty \quad \& x < 1 \\ 1 & n = \infty \quad \& x = 1 \end{cases} \quad (4.9)$$

is not a continuous function, since its restriction to $\{\infty\} \times [0, 1]$ is not continuous.

2. The compact-open topology is well-suited to discussing homotopy: A continuous path of functions in the compact-open topology is one way of defining a homotopy. To be a bit more precise: A continuous path

$$\wp : [0, 1] \rightarrow \mathcal{C}(X, Y) \quad (4.10)$$

which begins at a function $f_0 := \wp(0)$ and ends at $f_1 := \wp(1)$ is the same thing as a continuous map $F : [0, 1] \times X \rightarrow Y$ such that $F(0, x) = f_0(x)$ and $F(1, x) = f_1(x)$. That is the definition of homotopy discussed below.

3. One consequence of this, which is one of the primary sources of the applications of topology to field theory is the following: With the compact-open topology the set of homotopy classes of maps in $\mathcal{C}(X, Y)$, sometimes denoted $[X, Y]$, is the same as the set of disconnected components of the topological space $\mathcal{C}(X, Y)$ in the compact-open topology:

$$[X, Y] = \pi_0(\mathcal{C}(X, Y)) \quad (4.11)$$

4. Suppose Y is itself path-connected. Then there is a distinguished component of $\mathcal{C}(X, Y)$ containing any map taking all of X to a point.

Definition: A map homotopic to a constant map is called *null homotopic*

Exercise

- (a.) If (Y, d) is a metric space, show that the topology of compact convergence is finer than the topology of pointwise convergence.
 - (b.) Show that the compact-open topology is finer than the point-open topology.
-

5. Constructing topological spaces using gluing

A very common construction in physics involves making a quotient space by some kind of identification. Very generally we have an equivalence relation on a space X and we define $\tilde{X}$ to be the set of equivalence classes.

If X is a topological space then there is a natural way of making $\tilde{X} = X / \sim$ a topological space. We let $\pi : X \rightarrow \tilde{X}$ be the projection. This takes x to its equivalence class $[x] \in \tilde{X}$. We declare a set $U \subset \tilde{X}$ to be open iff $\pi^{-1}(U) \subset X$ is open. This defines what is known as the *quotient topology*. The quotient topology is the coarsest topology on $\tilde{X}$ so that the projection map π is continuous.

Examples

1. The space $\mathbb{RP}^2$ is defined as $S^2 / \sim$ where the equivalence relation identifies antipodal points of the sphere. If we try to represent each equivalence class by a single point then we can certainly throw away all the points in any open hemisphere. This makes it clear that the space can be identified as the space $D^2 / \sim$ where D^2 is the closed disk and now the equivalence relation identifies antipodal points (only on the boundary).
2. *Real projective space:* $\mathbb{RP}^n$ is the space obtained from the sphere S^n by the equivalence relation $x \sim -x$ identifying antipodal points.
3. *Real projective space:* $\mathbb{RP}^n$ is also the space of lines through the origin in $\mathbb{R}^{n+1}$. We can identify this as $\mathbb{R}^{n+1} - \{0\}$ with the identification $\vec{x} \sim \lambda \vec{x}$ for $\lambda \in \mathbb{R}^*$. In particular, the space $\mathbb{RP}^2$ is often called the *projective plane*.¹⁶
4. *Complex projective space:* The complex analog of the previous example is $\mathbb{CP}^n$, defined as the space of points in $\mathbb{C}^{n+1} - \{0\}$ with the identification $\vec{z} \sim \lambda \vec{z}$ for $\lambda \in \mathbb{C}^*$. A similar construction gives quaternionic projective space $\mathbb{HP}^n$.

¹⁶The equivalence classes $[x_1 : x_2 : x_3]$ can be put in the unique form $[x_1 : x_2 : 1]$ when $x_3 \neq 0$. These form a copy of $\mathbb{R}^2$, hence, the “plane.” However, the equivalence class $[x_1 : x_2 : 0]$ constitutes another “point at infinity.” Hence the name “projective plane.”

5. The “space of rays” in a Hilbert space are all the quantum states related to $\psi \in \mathcal{H}$ by $\psi \sim z\psi$ for $z \neq 0$. Much more on that in Chapter **** For finite-dimensional Hilbert spaces this is again complex projective space.
6. The gauge invariant configurations in Yang-Mills theory is the quotient space of the space of all connections with the equivalence relation of gauge transform.
7. All of the above examples are *quotients by group actions*, something we will study in more detail in Section §7 below. For an example of an equivalence relation which is not a group quotient consider the unit n -dimensional ball in $\mathbb{R}^n$:

$$D^n := \{\vec{x} \in \mathbb{R}^n \mid \vec{x} \cdot \vec{x} \leq 1\} \quad (5.1)$$

The interior of D^n is the set of points with $\vec{x} \cdot \vec{x} < 1$. We can quotient $D^n / \sim$ where the equivalence classes have just one element in the interior and we identify all of the points on the bounding sphere S^{n-1} to a single point. The quotient space with the quotient topology is homeomorphic to S^n .

A common way in which the quotient topology is used is if we are given a continuous map $f : X \rightarrow Y$ and a closed subspace $A \subset X$. Then we form the *glued space* $X \cup_f Y$ by identifying $a \in A$ with $f(a) \in Y$ for all $a \in A$. (The equivalence class is a singleton for all other points in $X \amalg Y$.)

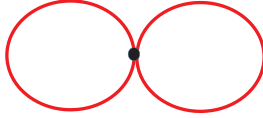


Figure 14: The one point union of two circles. The black point is the common identified point.

Examples

1. *One-point union.* If (X, x_0) and (Y, y_0) are *pointed spaces* - that is, spaces with a distinguished basepoint - then $X \vee Y$ is the pointed space by quotienting the disjoint union $X \amalg Y$ with the equivalence relation $x_0 \sim y_0$. (More formally, the equivalence class of points $x \neq x_0$ and $y \neq y_0$ have a single element and in addition there is one other equivalence class $\{x_0, y_0\}$.) Thus, for example, a figure 8 is a one-point union of two circles as in Figure 14. Here we take the closed subspace A of $X \amalg Y$ to be $A = \{x_0, y_0\}$, and $f : A \rightarrow \{p_0\}$ maps A to a single point.

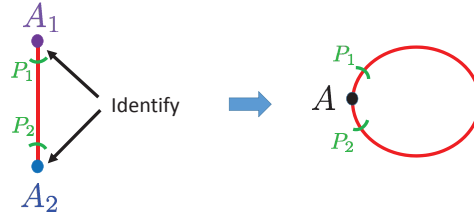


Figure 15: Identifying the endpoints A_1 and A_2 of the closed interval to a single point A we obtain the circle. The quotient topology is the standard topology of the circle. The inverse image of the interval between P_1 and P_2 on the circle is the union of *open* sets $[A_1, P_1)$ and $(P_2, A_2]$ in the interval $[A_1, A_2]$. Explain why the inverse image of the interval $[A, P_1)$ on the circle is not an open subset of the interval $[A_1, A_2]$.

2. Similarly, for the case of the disk D^n discussed above, $A = \partial D^n = S^{n-1}$ and we take $f : A \rightarrow \{p_0\}$ to map A to a single point, giving a space homeomorphic to S^n .
3. We can make a circle by identifying opposite ends of an interval as in Figure 15. We can similarly make an n -dimensional torus by identifying “opposite” sides of the n -dimensional cube $[0, 1]^n$ for any $n \geq 1$. See Figure 62 for the case $n = 2$.

6. Manifolds

Manifolds are topological spaces with some extra mathematical structure which makes them “everywhere locally like $\mathbb{R}^n$,” for some n . The heuristic idea is that, just as the earth appears locally flat, we understand that globally, the shape of the earth is more interesting. In general topological spaces can have very weird and unintuitive properties. These can be fun mathematically, but they are generally irrelevant in physical contexts, and thus, most of the topological spaces which are used in physics are manifolds. Moreover, the extra structure allows one to do calculus, which of course is essential to defining the differential equations, flows etc. that are fundamental to physics.

A few examples of the appearance of manifolds in physics are:

1. Space and spacetime are manifolds. The extra differentiable structure allows us to define fields and differential operators and equations on these fields. Moreover, global properties of the manifolds can be important both in quantum field theory and in general relativity. The very concept of general covariance is deeply entwined with the idea of manifolds. ¹⁷

¹⁷Successful tests of Lorentz invariance can be viewed as evidence that spacetime should be described as a pseudo-Riemannian manifold. The metric provides a length scale. Roughly speaking, experiments at energy scale E probe length scales of order hc/E . Numerically, we have approximately $hc \cong 1.2eV \cdot \mu m$. High

2. Phase spaces in Hamiltonian mechanics are manifolds with extra structure known as symplectic or Poisson structure. Classical dynamics is formulated in terms of flows determined by Hamiltonian vector fields. One approach to quantization stresses symplectic manifolds and becomes quite interesting and nontrivial for symplectic manifolds which are not $\mathbb{R}^{2n}$ with the standard Darboux symplectic form $\omega = dp_i dq^i$.
3. Spaces of particles are manifolds. Spaces of embeddings of worldlines of particles into spacetime, or of worldvolumes of branes into spacetime are interesting infinite-dimensional manifolds.
4. Spaces of solutions to differential equations are often manifolds. (Sometimes with singularities.) For example, moduli spaces of solutions to the Yang-Mills equations or to the holomorphic map equation are interesting manifolds.
5. Sometimes vacua in quantum field theory are not isolated but come with “moduli” these “moduli spaces of vacua” are often interesting manifolds.
6. Continuous symmetry groups are often Lie groups. By definition, Lie groups are manifolds where the group structure and the manifold structure are compatible. This extra structure leads to a very rich mathematical theory with a large number of applications to physics.

6.1 Basic Definitions

Here is the formal definition of a manifold:

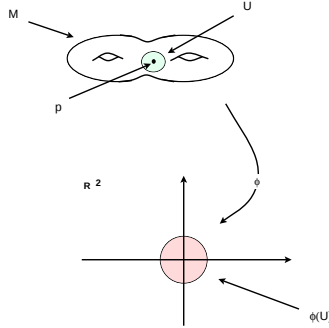


Figure 16: A coordinate patch on a manifold.

Definition : A ($\mathcal{C}^k$ differentiable) manifold M of dimension n is a topological space such that

1. There exists an open covering $\{U_\alpha\}_{\alpha \in I}$ of M , where I is an index set, together with homeomorphisms $\phi_\alpha : U_\alpha \rightarrow \mathbb{R}^n$.

energy collisions at the LHC provide tests down to a length scale of around $\frac{\hbar c}{E} \sim 10^{-18}m$ for $E = 1TeV$. The (probable) observation of the GZK cutoff on the spectrum of ultra-high-energy cosmic rays can be viewed as a test of special relativity to length scales $\frac{\hbar c}{E} \sim 10^{-26}m$. Many people suspect that at length scales given by the Planck length, $\ell_P := (\hbar G/c^3)^{1/2} \cong 1.6 \times 10^{-35}m$ the notion of a manifold structure of spacetime will have to be replaced by something else.

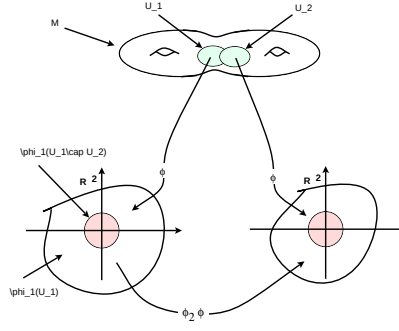


Figure 17: Transition functions between two coordinate patches. The map $\phi_2\phi_1^{-1}$ must be infinitely differentiable. FIX FIGURE!!

2. When $U_\alpha \cap U_\beta \neq \emptyset$ then the map

$$\phi_{\alpha\beta} := \phi_\alpha \circ \phi_\beta^{-1} : \phi_\beta(U_\alpha \cap U_\beta) \rightarrow \phi_\alpha(U_\alpha \cap U_\beta) \quad (6.1)$$

is a $\mathcal{C}^k$ map¹⁸ from $\phi_\beta(U_\alpha \cap U_\beta) \subset \mathbb{R}^n$ to $\phi_\alpha(U_\alpha \cap U_\beta) \subset \mathbb{R}^n$. See Figure 17.

We call the pair (U_α, ϕ_α) a *coordinate patch* or *coordinate chart*. See Figure 16. A collection of $\{(U_\alpha, \phi_\alpha)\}_{\alpha \in I}$ is called a *coordinate atlas*. In the case $k = 0$, the transition functions are continuous functions. A $\mathcal{C}^0$ -manifold is also known as a *topological manifold*. A $\mathcal{C}^\infty$ -manifold is also known as a *smooth manifold*. If $k > 0$ then associated with patch overlaps is a matrix of derivatives:

$$J[\phi_\beta; \phi_\alpha] \equiv \frac{\partial \phi_{\beta\alpha}^i(x)}{\partial x^j}$$

It is an $n \times n$ invertible matrix function of x defined on $\phi_1(U_1 \cap U_2)$.

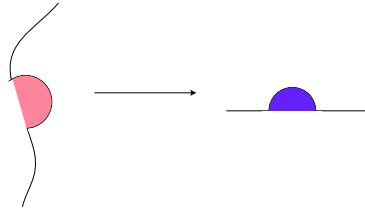


Figure 18: A coordinate patch of a manifold with boundary.

Definition : A *manifold with boundary* has local neighborhoods which look like $\mathbb{R}^n$ in the interior, and like the half space

$$H^n = \{(x_1, \dots, x_n) : x_n \geq 0\} \subset \mathbb{R}^n$$

on the boundary. See Figure 18.

¹⁸A $\mathcal{C}^k$ is one for which the first k derivatives are continuous.

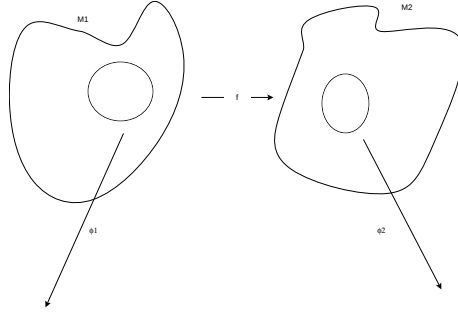


Figure 19: A $\mathcal{C}^k$ differentiable map between two manifolds. In all patches $\phi_2 f \phi_1^{-1}$ must be differentiable.

Definition : Let M_1, M_2 be $\mathcal{C}^k$ -manifolds with $k > 0$. Suppose $\dim M_i = n_i$.

a.) A map, or morphism, $f : M_1 \rightarrow M_2$ between two $\mathcal{C}^k$ - manifolds is called $\mathcal{C}^k$ differentiable if $\psi_\rho \circ f \circ \phi_\alpha^{-1}$ is a $\mathcal{C}^k$ differentiable map from an open set of $\mathbb{R}^{n_1}$ to an open set of $\mathbb{R}^{n_2}$ for all coordinate atlases $\{U_\alpha, \phi_\alpha\}$ of M_1 and $\{V_\rho, \psi_\rho\}$ of M_2 .

b.) In particular, a function $f : M_1 \rightarrow \mathbb{R}$ is differentiable at p if there is a coordinate chart containing p so that $f \circ \phi^{-1} : \mathbb{R}^n \rightarrow \mathbb{R}$ is $\mathcal{C}^k$ -differentiable.

c.) A map $f : M_1 \rightarrow M_2$ is a *diffeomorphism* if f is an infinitely differentiable homeomorphism with an infinitely differentiable inverse.

d.) The set of automorphisms of a smooth manifold M is a group known as the *diffeomorphism group of M* , and denoted $\text{Diff}(M)$.

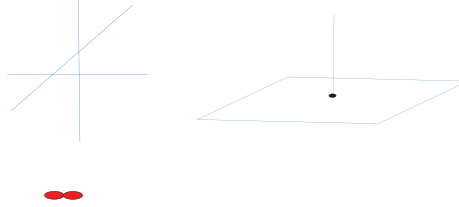


Figure 20: Some topological spaces which are not manifolds.

Remarks.

1. Some examples of topological spaces which are not manifolds are shown in Figure 20. See also Figure 108 below.
2. There are many variations on the above definitions. For example one can define *manifolds with corners*. The coordinate charts are now modelled on $\mathbb{R}^n$ or on

$$\mathbb{R}_+^n = \mathbb{R}_+^k \times \mathbb{R}^{n-k} \quad (6.2)$$

where $\mathbb{R}_+ = [0, \infty)$ where some charts have $k > 1$. If all $k \leq 1$ we have a manifold with boundary. Thus, for example we distinguish $\mathbb{R}_+^2$ from $\mathbb{R}_+ \times \mathbb{R}$. The first can be identified with a standard quadrant in the plane and the second with a half-plane. If we think of the plane as the complex plane then the two are homeomorphic by the transformation $z \mapsto z^2$. However, the inverse is not differentiable at $z = 0$. Thus, the distinction between manifolds with corners starts to make sense at the level of diffeomorphism class. For more detail see. ¹⁹

3. Fussbudgets will find reasons to criticize our definition. One should say that M is a paracompact Hausdorff space. (Some authors do not require M to be Hausdorff.) Moreover, one should define a maximal atlas as one that is not a proper subset of any other atlas, and, strictly speaking a manifold is a paracompact ²⁰ Hausdorff space equipped with a maximal atlas. Typical examples in physics will be paracompact and niceties of choosing a maximal atlas can, in general, be safely ignored. ²¹
4. There is an algebraic approach to manifolds. Note that $\mathcal{C}^k(M)$, the vector space of all real-valued k -fold differentiable functions on M is in fact an *algebra*: ²² The basic operations are
 - a.) Scalar multiplication: $(\alpha \cdot f)(p) := \alpha f(p)$.
 - b.) Vector addition: $(f + g)(p) := f(p) + g(p)$.
 - c.) Algebra multiplication: $(f \cdot g)(p) := f(p)g(p)$.

One instantly checks the axioms. In fact, as a topological space, (i.e. the homeomorphism class), M is completely determined by the algebra $\mathcal{C}^0(M)$. (This is known as Gelfand's theorem.) This algebraic viewpoint is useful for various generalizations of the notion of a manifold, both in algebraic geometry and in noncommutative geometry. We will find it a useful viewpoint when defining the tangent and cotangent spaces below.

5. It is very important to understand the covariance and contravariance properties of structures associated with manifolds. For the most elementary example suppose $f : M_1 \rightarrow M_2$ is a $\mathcal{C}^k$ morphism of $\mathcal{C}^k$ manifolds. Then the *pull back by f* maps functions on M_2 to functions on M_1 by

$$\begin{aligned} f^* : \mathcal{C}^k(M_2) &\rightarrow \mathcal{C}^k(M_1) \\ F &\mapsto f^*(F) := F \circ f \end{aligned} \tag{6.3}$$

Note this is a morphism of algebras:

¹⁹Ref: D. Joyce, "On manifolds with corners," <http://arxiv.org/pdf/0910.3518v2.pdf>

²⁰A cover is *locally finite* if every point p has a neighborhood U with $U \cap U_\alpha \neq \emptyset$ for only finitely many α . A *refinement* of a cover $\{U_\alpha\}$ is a cover $\{V_\alpha\}$ so that $V_\alpha \subset U_\alpha$. A topological space is *paracompact* if every open cover has a locally finite refinement. The importance of this notion is that it is used to define partitions of unity.

²¹For example, a search on google with [site:arxiv.org manifold] retrieves about 181,000 results while a search with [site:arxiv.org "maximal atlas"] retrieves 277 results.

²²See Chapter *** for a systematic discussion of algebras.

- a.) $f^*(\alpha F) = \alpha f^*(F)$
 b.) $f^*(F + G) = f^*(F) + f^*(G)$
 c.) $f^*(FG) = f^*(F)f^*(G)$
6. It is also possible, and important, to define infinite-dimensional manifolds. The local coordinate charts must have extra structure to have a good theory, thus they will be Banach spaces, or Hilbert spaces, for examples. If M, N are manifolds then we can consider the space of maps $\text{Map}(M, N)$ to be a manifold provided we suitably restrict the set of functions. If M, N are both of positive dimension then $\text{Map}(M, N)$ is infinite-dimensional. In field theory, spaces of field configurations are typically infinite-dimensional manifolds.

Exercise

Show that

$$\phi_{\alpha\beta} \circ \phi_{\beta\alpha} = Id \quad (6.4)$$

$$\phi_{\alpha\beta} \circ \phi_{\beta\gamma} \circ \phi_{\gamma\alpha} = Id \quad (6.5)$$

where Id is the identity map, and describe carefully the domains on which these equations make sense.

Exercise

If M_1, M_2 are manifolds then $M_1 \times M_2$ is a manifold.

Exercise *Connected Sum*

The following operation on two manifolds M_1, M_2 of the same dimension is called the *connected sum*. Choose points $p_i \in M_i$ and remove a small ball B_i from around each point. Now $M_i - B_i$ is a manifold with boundary given by the sphere ∂B_i . Glue the two spheres together to produce a new manifold $M_1 \# M_2$.

- a.) What is the dimension of $M_1 \# M_2$?
 b.) If M_1 is of dimension n show that $M_1 \# S^n \cong M_1$.
 c.) Show that a connected sum of a torus and $\mathbb{R}P^2$ is equivalent to a connected sum of three copies of $\mathbb{R}P^2$.

Remark: A beautiful and nontrivial theorem says that under connected sum all orientable 3-manifolds have a unique “prime decomposition.” Reference: J. Milnor, A unique decomposition theorem for 3-manifolds, American Journal of Mathematics 84 (1962), 17.

6.2 Examples

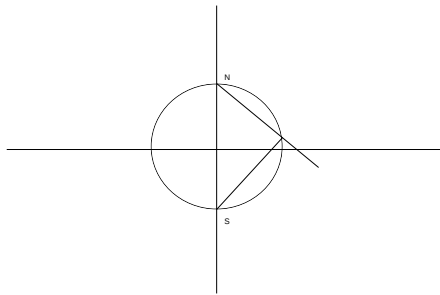


Figure 21: Stereographic projection.

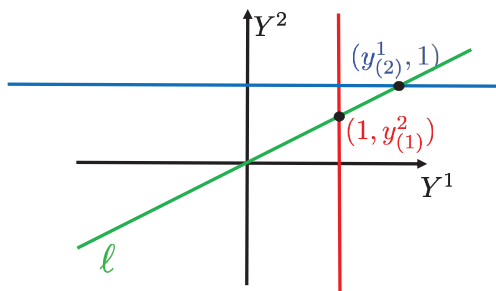


Figure 22: Illustrating coordinates and charts for $\mathbb{RP}^1$. The patch U_1 consists of all the green lines through the origin of $\mathbb{R}^2$ except for the (unoriented!) line $Y^1 = 0$. The affine hyperplanes H_1 and H_2 defined by $Y^1 = 1$ and $Y^2 = 1$, respectively, are illustrated. The homeomorphism ϕ_1 from U_1 to H_1 is determined by the intersection point $(1, y_{(1)}^2)$, and similarly for ϕ_2 . Note well that $y_{(1)}^2 y_{(2)}^1 = 1$.

1. *One-dimensional manifolds.* Up to diffeomorphism there are exactly two connected one-dimensional manifolds without boundary. These are $\mathbb{R}$ itself and S^1 . $\mathbb{R}$ is diffeomorphic to (a, b) for any pair of real numbers $a < b$. For example the map diffeomorphism

$$x \rightarrow \frac{1}{1 + e^{-x}}$$

provides a diffeomorphism $\mathbb{R} \cong (0, 1)$. Of course, since $\mathbb{R}$ has infinitely many self-diffeomorphisms there are infinitely many such diffeomorphisms. If there is one boundary then the one-dimensional manifold is diffeomorphic to $[0, \infty)$, which is in turn diffeomorphic to $[a, b)$ for any pair of real numbers $a < b$. If there are two boundaries then M is the compact 1-manifold $[0, 1]$. It is diffeomorphic to $[a, b]$ for any pair of real numbers $a < b$. By taking disjoint unions with these connected manifolds one can make the general one-manifold.

2. *Spheres.* Some manifolds are defined as subspaces of Euclidean space satisfying certain equations. A good example are the spheres $S^n \subset \mathbb{R}^{n+1}$ defined as the solution set of $f(\vec{x}) = \vec{x}^2 - 1 = 0$. The sphere is not homeomorphic to $\mathbb{R}^n$ so any description as a manifold requires at least two coordinate patches. In fact, two patches will suffice. One very natural choice is provided by stereographic projection as in Figure 21.

Coordinate patches:

$$\mathcal{U}_- = S^n - \{\wp_N = (\vec{0}, 1)\}$$

$$\mathcal{U}_+ = S^n - \{\wp_S = (\vec{0}, -1)\}$$

Local homeomorphisms: $\phi_{\pm}$ are just given by stereographic projection

$$\phi_-(\vec{x}) = \frac{\vec{x}}{1 - x_{n+1}} = \vec{y}_S$$

$$\phi_+(\vec{x}) = \frac{\vec{x}}{1 + x_{n+1}} = \vec{y}_N$$

Overlap: Note that $\phi_{\pm}(\mathcal{U}_- \cap \mathcal{U}_+) = \mathbb{R}^n - \{0\}$. To compute transition functions note that:

$$\vec{y}_S^2 = \frac{1}{\vec{y}_N^2} = \frac{1 + x_{n+1}}{1 - x_{n+1}} \quad (6.6)$$

so the transition functions are

$$\vec{y}_S = \frac{1 + x_{n+1}}{1 - x_{n+1}} \vec{y}_N = \frac{\vec{y}_N}{\vec{y}_N^2} \quad (6.7)$$

Or, equivalently,

$$\vec{y}_N = \frac{1 - x_{n+1}}{1 + x_{n+1}} \vec{y}_S = \frac{\vec{y}_S}{\vec{y}_S^2} \quad (6.8)$$

3. *Real and complex projective space.* Another way to define manifolds is via a quotient construction using an equivalence relation. We defined above the topological spaces $\mathbb{FP}^n$ for $\mathbb{F} = \mathbb{R}, \mathbb{C}$ (and it also works for $\mathbb{F} = \mathbb{H}$ if we are careful about which side we use to multiply by “scalars” in $\mathbb{H}$). These were defined above as topological spaces by putting an equivalence relation on $\mathbb{F}^{n+1} - \{0\}$ with an equivalence class given by

$$[\vec{x}] = \{\lambda \vec{x} \in \mathbb{F}^{n+1} | \lambda \in \mathbb{F}^*\} \quad (6.9)$$

Note that such an equivalence class uniquely determines a one-dimensional subspace in $\mathbb{F}^{n+1}$:

$$\ell_{\vec{x}} = \{\lambda \vec{x} \in \mathbb{F}^{n+1} | \lambda \in \mathbb{F}\} \quad (6.10)$$

(A one-dimensional subspace of a vector space is often called a *line*.) The only difference between (6.9) and (6.10) is that in the latter we allow $\lambda = 0$, as is necessary to have a linear subspace. Conversely, any one-dimensional subspace of $\mathbb{F}^{n+1}$ is of the form (6.10) and this uniquely determines the class (6.9). *Thus, we can view $\mathbb{FP}^n$*

as the moduli space of lines in $\mathbb{F}^{n+1}$. Another, very useful way of denoting the equivalence class is the notation

$$[X^1 : X^2 : \dots : X^{n+1}] = [\lambda X^1 : \lambda X^2 : \dots : \lambda X^{n+1}] \quad \lambda \neq 0 \quad (6.11)$$

where $\vec{X} = (X^1, X^2, \dots, X^{n+1}) \in \mathbb{F}^{n+1} - \{0\}$. The X^i in (6.11) are called “homogeneous coordinates” although they are not really coordinates since there is no definite value we can assign to X^i if we are just given a point in $\mathbb{F}\mathbb{P}^n$. To get true coordinates we need coordinate patches. For any $1 \leq \alpha \leq n+1$ define

$$U_\alpha := \{[X^1 : X^2 : \dots : X^{n+1}] | X^\alpha \neq 0\} \quad (6.12)$$

We can “fix the gauge” on U_α , that is, fix the ambiguity in the values of the X^i by rescaling by λ by choosing the representative in the equivalence class $[Y^1 : Y^2 : \dots : Y^{n+1}]$ with $Y^\alpha = 1$. For this unique representative $\vec{Y} \in H_\alpha \subset \mathbb{F}^{n+1}$ where H_α is the affine hyperplane defined by the equation $Y^\alpha = 1$. Clearly, we could choose coordinates on H_α using vectors in $\mathbb{F}^n$. Therefore the map

$$\begin{aligned} \phi_\alpha^{-1} : H_\alpha &\rightarrow U_\alpha \\ \phi_\alpha^{-1} : \vec{Y} &\mapsto [Y^1 : \dots : Y^{n+1}] \end{aligned} \quad (6.13)$$

provides a local homeomorphism of $\mathbb{F}^n$ with U_α . Thus, the collection $\{(U_\alpha, \phi_\alpha)\}_{1 \leq \alpha \leq n+1}$ forms a coordinate atlas for $\mathbb{F}\mathbb{P}^n$.

Example 1: The coordinate system for $\mathbb{R}\mathbb{P}^1$ is illustrated in Figure 22.

Example 2: Choose $n = 2$ then

$$\begin{aligned} \varphi_1^{-1} : (1, Y^2, Y^3) &\mapsto [1 : Y^2 : Y^3] \\ \varphi_2^{-1} : (Y^1, 1, Y^3) &\mapsto [Y^1 : 1 : Y^3] \\ \varphi_3^{-1} : (Y^1, Y^2, 1) &\mapsto [Y^1 : Y^2 : 1] \end{aligned} \quad (6.14)$$

Now let us compute the transition functions. Note that $\phi_\beta(U_\alpha \cap U_\beta)$ is the subspace of H_β of vectors with $Y^\alpha \neq 0$. Now, for $\alpha \neq \beta$ to compute $\phi_\alpha \circ \phi_\beta^{-1}(\vec{Y})$ we note that

$$\phi_\beta^{-1}(\vec{Y}) = [Y^1 : \dots : Y^{n+1}] \quad (6.15)$$

and $\vec{Y} \in \phi_\beta(U_\alpha \cap U_\beta)$ means that we have both $Y^\beta = 1$ and $Y^\alpha \neq 0$. In order to compute the image under Y^α we must choose the representative of the equivalence class with $\tilde{Y}^\alpha = 1$ and hence we should apply (6.11) with $\lambda = 1/Y^\alpha$:

$$\phi_\beta^{-1}(\vec{Y}) = [Y^1/Y^\alpha : \dots : Y^{n+1}/Y^\alpha] \quad (6.16)$$

Therefore

$$\phi_\alpha \circ \phi_\beta^{-1} : H_\beta - \{Y^\alpha \neq 0\} \rightarrow H_\alpha - \{Y^\beta \neq 0\} \quad (6.17)$$

is simply given by

$$\phi_\alpha \circ \phi_\beta^{-1}(\vec{Y}) = \frac{1}{Y^\alpha} \vec{Y} \quad (6.18)$$

This is clearly differentiable. Moreover, for $\mathbb{F} = \mathbb{C}$ the transition maps are holomorphic. This leads to the important concept of a *complex manifold*. These are manifolds where the coordinate patches can be identified with $\mathbb{C}^n$ and the transition functions are determined by n holomorphic functions of n complex variables.

4. *Grassmannians*. Grassmannians generalize the projective spaces. The projective spaces are the spaces of all lines in a vector space. The Grassmannian $\text{Gr}_k(\mathbb{F}^n)$ is the space of all k -dimensional subspaces of $\mathbb{F}^n$. We defer a detailed discussion of this manifold until later.

Exercise

Any complex manifold is *a fortiori* also a real manifold.

- What is the dimension of a complex manifold of dimension n as a real manifold?
 - Show that, regarded as a real manifold $\mathbb{CP}^1$ is diffeomorphic to S^2 .²³
-

Exercise *Moduli space of hyperplanes*

Consider the set of all n -dimensional linear subspaces of the vector space $\mathbb{F}^{n+1}$, where $\mathbb{F} = \mathbb{R}, \mathbb{C}$. (Two separate cases.)

- Show that these manifolds are diffeomorphic to certain manifolds discussed in the notes.
 - What is the dimension of these manifolds? (Caution: Trick question.)
-

Exercise

Is $\mathbb{CP}^n$ compact? Is $\text{Gr}_k(\mathbb{C}^n)$ compact?

Exercise

²³ *Answer:* For $\mathbb{CP}^1$ we have $\phi_1 : [X^1 : X^2] \rightarrow (1, Y_{(1)}^2)$ on the patch U_1 with $X^1 \neq 0$ and similarly $\phi_2 : [X^1 : X^2] \rightarrow (Y_{(2)}^1, 1)$ on the patch U_2 with $X^2 \neq 0$. So we can write $\phi_1 \circ \phi_2^{-1} : (z, 1) \mapsto (1, 1/z)$ where $z \neq 0$. In other words we have two patches with coordinates $z_{\pm} \in \mathbb{C}$ and on the overlap $\mathbb{C}^*$ we have $z_+ = 1/z_-$. Relate this to the two-dimensional sphere via stereographic projection (with complex conjugation on one patch):

$$z_+ = \frac{x^1 + ix^2}{1 - x^3} \quad z_- = \frac{x^1 - ix^2}{1 + x^3} \quad (6.19)$$

One way of defining a coordinate patch around any line $\ell \subset \mathbb{F}^{n+1}$ is the following. Choose the standard inner product on $\mathbb{F}^{n+1}$, and denote the inner product of v, w by (v, w) . (For $\mathbb{F} = \mathbb{C}$ it is antilinear on the first variable.) If ℓ is a line then there is a uniquely determined orthogonal subspace $\ell^\perp \subset \mathbb{F}^{n+1}$. Let $T \in \text{Hom}(\ell, \ell^\perp)$ be any $\mathbb{F}$ -linear transformation. Then T determines a line given by the graph of T :

$$L_{\ell, T} := \{v + T(v) | v \in \ell\} \subset \mathbb{F}^{n+1} \quad (6.20)$$

Then the canonical coordinate patch around ℓ is

$$U_\ell = \{L_{\ell, T} | T \in \text{Hom}(\ell, \ell^\perp)\} \quad (6.21)$$

a.) Show that $\text{Hom}(\ell, \ell^\perp) \cong \mathbb{F}^n$ and describe the coordinate functions.

b.) For two distinct lines ℓ_1, ℓ_2 let P_2 denote the orthogonal projection operator to ℓ_2 . Show that $L_{\ell_1, T_1} = L_{\ell_2, T_2}$ iff for all $v \in \ell_1$,

$$T_2 \circ P_2(v + T_1(v)) = (1 - P_2)(v + T_1(v)) \quad (6.22)$$

and this completely determines T_2 in terms of ℓ_1, ℓ_2, T_1 . Using this compute the transition functions.

6.3 Tangent and cotangent space

In order to do calculus on manifolds we need the notion of tangent and cotangent vectors. One of the key properties of manifolds, as opposed to general topological spaces is that at every point p of a manifold M there are two canonically associated vector spaces called the *tangent space* $T_p M$ and the *cotangent space* $T_p^* M$. Moreover, these vector spaces are canonically dual vector spaces.

The tangent space $T_p M$ generalizes our intuitive notion of tangent vectors to a curve in the plane, or to a surface in three-dimensional space.

Locally, near p , the manifold M “looks like” $\mathbb{R}^n$. Now $\mathbb{R}^n$ has the extra mathematical structure of a vector space, so we might simply try to associate a tangent vector to a vector $\vec{v} \in \mathbb{R}^n$. The problem is that the linear structure on $\mathbb{R}^n$ is not preserved under change of coordinates. Closely related to this is that the gluing functions $\phi_{\alpha\beta}$ are in general not affine linear transformations.

6.3.1 Definition of $T_p M$ using directional derivatives

To introduce the general idea suppose $(a, b) \subset \mathbb{R}$ is an open interval containing 0 and $\lambda : (a, b) \rightarrow \mathbb{R}^n$ is a differentiable curve passing through $\lambda(0) = p$. The curve is described by an n -tuple of functions

$$\lambda(t) = (\lambda^1(t), \dots, \lambda^n(t)) \quad (6.23)$$

and the tangent vector at p is simply:

$$\begin{aligned} \vec{v}(p) &= \sum_{i=1}^n v^i(p) \vec{e}_i := v^i(p) \vec{e}_i \\ v^i(p) &:= \frac{d\lambda^i}{dt} \Big|_{t=0} \end{aligned} \quad (6.24)$$

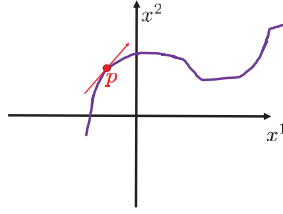


Figure 23:

where we use the summation convention in the first line. See Figure 23.

Now suppose that $p \in U \subset M$ and $\lambda : (a, b) \rightarrow M$ is a differentiable path in M with $\lambda(0) = p$. If we choose coordinates $\phi : U \rightarrow \mathbb{R}^n$ with coordinate functions $\phi(q) = (x^1(q), \dots, x^n(q))$, $q \in U$, and then $\phi \circ \lambda$ is a path in $\mathbb{R}^n$ with

$$\lambda^i(t) := x^i(\lambda(t)). \quad (6.25)$$

Again we can associate a vector $\vec{v}(p)$ as above.

The problem with this is that if we use some other coordinate system $\tilde{\phi}$ then if $\tilde{\phi} \circ \phi^{-1}$ is described by the general differentiable change of coordinates

$$y^i = y^i(x^1, \dots, x^n) \quad (6.26)$$

we get a new vector with coefficients

$$\tilde{v}^i(p) = \frac{d\tilde{\lambda}^i}{dt} \Big|_{t=0} = \frac{\partial y^i}{\partial x^j} \Big|_p v^j(p) \quad (6.27)$$

where $\tilde{\lambda}^i = y^i(\lambda^1(t), \dots, \lambda^n(t))$. One could define tangent vectors as n -tuples that transform under coordinate transformations in this way, but there is a better, more invariant way to proceed.

To motivate it, return to our path $\lambda : (a, b) \rightarrow \mathbb{R}^n$ through p . Associated to λ is a *directional derivative* at p . Suppose f is a C^1 function “defined near p .” This means that $f \in C^1(U)$ where U is some open neighborhood of p in $\mathbb{R}^n$. Then the directional derivative of f through λ at p is defined to be:

$$\frac{d}{dt} \Big|_{t=0} f(\lambda(t)) \quad (6.28)$$

Now (6.28) assigns to any function a real number. Moreover the map ²⁴

$$\begin{aligned} \lambda_* \left(\frac{d}{dt} \right) : C^1(U) &\rightarrow \mathbb{R} \\ \lambda_* \left(\frac{d}{dt} \right) : f &\mapsto \frac{d}{dt} \Big|_{t=0} f(\lambda(t)) \end{aligned} \quad (6.29)$$

defines a linear functional on the vector space $C^1(U)$.

Remarks:

²⁴The reason for the strange notation is given in section **** below.

1. From the explicit expression (6.28) it clearly encodes the information in the traditional vector field (6.24). Indeed, we could let $f = x^i$ to obtain the individual $v^i(p)$.
2. The linear operator depends very weakly on U . The linear operator depends only on the “local” behavior of f . Two real-analytic functions have the same derivatives at p will certainly have the same value. So, we can refine our operator to a linear operator

$$C^1(p) \rightarrow \mathbb{R} \quad (6.30)$$

where $C^1(p)$ is the vector space of functions which are differentiable at p - i.e. in *some* neighborhood of p .

3. The operator $\lambda_*(\frac{d}{dt})$ also depends very weakly on the choice of the path $\lambda(t)$. There are infinitely many paths through p but if two paths have the same values of $\frac{d\lambda^i(t)}{dt}|_{t=0}$ then they define the same directional derivative. This puts an equivalence relation on the space of differentiable paths in $\mathbb{R}^n$ through p .
4. Note that the family of linear operators $\lambda_*(\frac{d}{dt})$ is in fact itself a vector space. The equivalence class of any operator $\lambda_*(\frac{d}{dt})$ has a unique representative curve of the form

$$\lambda^i(t) = x_0^i + tv^i \quad (6.31)$$

for $\vec{v} \in \mathbb{R}^n$. The usual vector space structure on v^i defines one on the space of directional derivatives.

The important aspect of this point of view is that it immediately generalizes to an arbitrary differentiable manifold: Note that *without choosing a coordinate system*, given a path λ , the linear functional

$$f \mapsto \frac{d}{dt}|_{t=0} f(\lambda(t)) \quad (6.32)$$

makes sense. If we choose a coordinate system ϕ we can describe the linear functional in terms of paths in $\mathbb{R}^n$ as above. If we have two different coordinate systems the two linear spaces are related by the linear transformation (6.27).

Definition 1 Let M be a differentiable manifold and $p \in M$. The *tangent space to M at p* is the vector space of linear operators

$$C^1(p) \rightarrow \mathbb{R} \quad (6.33)$$

defined by:

$$T_p M = \left\{ \lambda_*\left(\frac{d}{dt}\right) : \mid \lambda \text{ differentiable curve with } \lambda(0) = p \right\} \quad (6.34)$$

To spell out the isomorphism $T_p M \cong \mathbb{R}^n$ provided by a coordinate system, choose local coordinates x^i and fix an index i_0 . Consider the curve defined by

$$\lambda_{i_0}^i(t) = x_0^i + t\delta_{i_0}^i x^{i_0}. \quad (6.35)$$

Then the corresponding directional derivative is

$$(\lambda_{i_0})_* \left(\frac{d}{dt} \right) (f) = \frac{\partial f \circ \phi^{-1}}{\partial x^{i_0}}|_p \quad (6.36)$$

We call this operator $\frac{\partial}{\partial x^i}|_p$. Clearly

$$(\lambda)_* \left(\frac{d}{dt} \right) = v^i(p) \frac{\partial}{\partial x^i}|_p \quad (6.37)$$

Thus, these vectors span $T_p M$. They are also linearly independent since they have different values on the functions x^i , and hence form a basis for the vector space $T_p M$. Such a basis, determined by a coordinate chart, is known as a *coordinate basis*.

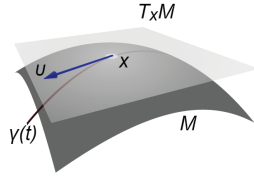


Figure 24: From commons.wikipedia.org. Need to change notation.

Example 1: Suppose $M \subset \mathbb{R}^n$ is a subset defined by $f(\vec{x}) = 0$ and suppose $\vec{x}_0 \in M$ has $\nabla f(\vec{x}_0) \neq 0$. Such a point is called a *regular point*. We will see below that this is the condition for f to define a manifold in the neighborhood of $\vec{x}_0$. By differentiating the equation $f(\lambda(t)) = 0$ for a path λ on M one easily sees that the tangent space $T_{\vec{x}_0} M$ can be directly identified with the linear space:

$$T_{\vec{x}_0} M \cong \left\{ \vec{w} \in \mathbb{R}^n \mid \vec{w} \cdot \vec{\nabla} f(\vec{x}_0) = 0 \right\} \subset \mathbb{R}^n \quad (6.38)$$

So our notion of a tangent space is in accord with what one would intuitively call a tangent space. See Figure 24. Note particularly that

1. There exists a vector space $E_{\vec{x}_0}^\perp$ such that there is an orthogonal decomposition:

$$T_{\vec{x}_0} M \oplus E_{\vec{x}_0}^\perp = \mathbb{R}^n \quad (6.39)$$

Indeed we can take $E_{\vec{x}_0}^\perp$ to be the linear span of the vector $\vec{\nabla} f(\vec{x}_0)$ in $\mathbb{R}^n$.

2. Therefore, the tangent space can be defined as the image of an orthogonal projection operator. When we vary the point $\vec{x}_0 \in M$ we get a (continuous, or differentiable) family of orthogonal projection operators.

These properties are actually quite generally true for manifolds, as one can prove rather easily using a partition of unity.

Example 2: *Tangent space to projective space.* Suppose ℓ_t is a path in $\mathbb{FP}^n$ passing through some line ℓ_0 . We may describe it in homogeneous coordinates $[X^1(t) : \dots : X^{n+1}(t)]$. By making a linear change of the homogeneous variables we may suppose that $X^1(0) \neq 0$ while $X^i(0) = 0$ for $i > 1$. Then $X^1(t) \neq 0$ for t near zero, so we can describe the curve as

$$[1 : Y^2(t) : \dots : Y^{n+1}(t)] \quad (6.40)$$

and the coordinates under ϕ_1 are just $(1, Y^2(t), \dots, Y^{n+1}(t)) \in H_1 \subset \mathbb{F}^{n+1}$. The derivative gives the vector $(0, \dot{Y}^2, \dots, \dot{Y}^{n+1})$ where $\dot{Y}^i := \frac{dY^i(t)}{dt}|_0$. Note that if we give $\mathbb{F}^{n+1}$ its standard inner product then this is an orthogonal vector to the vector $(1, 0, \dots, 0)$. We can therefore associate to this tangent vector an element of $\text{Hom}(\ell, \ell^\perp)$. It can be shown that this association is independent of all choices and is suitably invariant under the action of $GL(n+1, \mathbb{F})$ on $\mathbb{FP}^n$. Therefore, the invariant description of the tangent space at $\ell \in \mathbb{FP}^n$ is that it is the vector space

$$T_\ell \mathbb{FP}^n = \text{Hom}(\ell, \ell^\perp) \quad (6.41)$$

Example 3: *Tangent space to field space.* In the *nonlinear sigma model* we consider a theory of fields $\Psi : \mathcal{S} \rightarrow M$ from some space, or spacetime manifold $\mathcal{S}$ to a target manifold M . This target space can be simply a finite-dimensional vector space of dimension N , in which case we are simply studying a theory of N scalar fields. The space of fields is then $\text{Map}(\mathcal{S}, M)$ where, in a rigorous treatment we would need to put some conditions on the fields.²⁵ Let us simply take the differentiable fields. Let us compute the tangent space at some field configuration $\Psi_0 \in \text{Map}(\mathcal{S}, M)$. Let Ψ_t be a family of field configurations passing through Ψ_0 at time $t = 0$. We assume that $t \mapsto \Psi_t$ is differentiable at $t = 0$ once one puts a suitable differentiable structure on the infinite-dimensional manifold $\text{Map}(\mathcal{S}, M)$. Then use:

$$\text{Map}(\mathbb{R}, \text{Map}(\mathcal{S}, M)) \cong \text{Map}(\mathbb{R} \times \mathcal{S}, M) \cong \text{Map}(\mathcal{S}, \text{Map}(\mathbb{R}, M)) \quad (6.42)$$

Now an element of $\text{Map}(\mathbb{R}, M)$ is just a path in M . Thus, if $\sigma \in \mathcal{S}$ is a point in the domain spacetime we can fix σ and consider the path $\lambda(t; \sigma) := \Psi_t(\sigma)$ as a path in M . This has a tangent vector

$$\Phi(\sigma) := \lambda(\cdot; \sigma)_* \left(\frac{d}{dt} \right) |_{t=0} \in T_{\Psi_0(\sigma)} M \quad (6.43)$$

If all goes smoothly then Φ is a smooth map from $\mathcal{S}$ to tangent vectors on M taking σ into a tangent vector in the vector space $T_{\Psi_0(\sigma)} M$. Thus Φ is another kind of field. The space of such fields Φ is called the space of *sections of the pullback of the tangent bundle*, and so the tangent space is

$$T_{\Psi_0} \text{Map}(\mathcal{S}, M) \cong \Gamma(\Psi_0^*(TM)) \quad (6.44)$$

²⁵We are being deliberately vague here since a careful treatment would entail many details. Roughly speaking we want fields such that actions and energies are well-defined.

See the chapter **** below on the theory of bundles for explanation of the notation.

Exercise

Suppose $\{x^i\}, \{y^i\}$ are two coordinate systems in a neighborhood $p \in M$. Show that the coordinate bases for $T_p M$ are related by

$$\frac{\partial}{\partial x^i} = \frac{\partial y^j}{\partial x^i} \frac{\partial}{\partial y^j} \quad (6.45)$$

6.3.2 Definition of $T_p M$ in terms of derivations

Let $\mathcal{C}^\infty(U)$ denote the vector space of real differentiable functions on an open subset U of M . This is an algebra. We now recall a bit of abstract algebra:

Definition Let $\mathcal{A}$ be an algebra over a field (or ring) κ and $\mathcal{M}$ a (left) module for $\mathcal{A}$. A *derivation* D of $\mathcal{A}$ is a linear map:

$$D : \mathcal{A} \rightarrow \mathcal{M} \quad (6.46)$$

satisfying the Leibniz rule. That is it satisfies the axioms:

Der1: $D(\alpha f + \beta g) = \alpha Df + \beta Dg \quad \forall \alpha, \beta \in \kappa, \forall f, g \in \mathcal{A}$.

Der2: $D(fg) = fDg + gDf$

We will be concerned with $\mathcal{A}$ the algebra of real differentiable functions on U or M . Hence the field $\kappa = \mathbb{R}$. We first take the module $\mathcal{M}_p$ defined by a point p . As a vector space $\mathcal{M}_p \cong \mathbb{R}$ and f acts on $\mathcal{M}_p$ by multiplication by $f(p)$. Then a derivation is just a linear map

$$\mathcal{V}_p : \mathcal{C}^\infty(M; \mathbb{R}) \rightarrow \mathcal{M}_p \cong \mathbb{R} \quad (6.47)$$

which satisfies the Leibniz rule:

$$\mathcal{V}_p(fg) = f(p)\mathcal{V}_p(g) + \mathcal{V}_p(f)g(p) \quad (6.48)$$

Using this concept we can give our second definition of the tangent space $T_p M$:

Definition 2 The *tangent space* $T_p M$ to M at p is the vector space of derivations at p .

Why is this equivalent to the previous definition? Suppose (U, x^i) is a coordinate chart. Certainly $\frac{\partial}{\partial x^i}|_p$ define derivations on the algebra of differentiable functions in a neighborhood of p . Moreover the functions x^i on U generate $\mathcal{C}^\infty(U; \mathbb{R})$.²⁶ So, if $\mathcal{V}_p$ is a derivation we can define:

$$v^i(p) := \mathcal{V}_p(x^i) \quad (6.49)$$

²⁶ Actually, algebraically they generate the subspace of polynomial functions. We can complete that in various ways, so this is enough.

and then

$$\mathcal{V}_p = \sum_i v^i(p) \frac{\partial}{\partial x^i} |_p \quad (6.50)$$

We can now consider families of vectors for p in an open set U to define *vector fields* on U . We now consider the module $\mathcal{M} = \mathcal{A}$ to be the space of functions on U itself.

Definition A real $\mathcal{C}^\infty$ *vector field* on U is a derivation of the algebra $\mathcal{A} = \mathcal{C}^\infty(U; \mathbb{R})$ valued in $\mathcal{A}$. We denote the space of vector fields on U by $\text{Der}(U)$.

Exercise *The bracket of two vector fields*

Let us compute the bracket. Suppose that in a local coordinate chart (U, x^i) we have two vector fields:

$$\begin{aligned} \mathcal{V}_1 &= \sum_i a^i(x) \frac{\partial}{\partial x^i} \\ \mathcal{V}_2 &= \sum_i b^i(x) \frac{\partial}{\partial x^i} \end{aligned} \quad (6.51)$$

a.) Show that:

$$[\mathcal{V}_1, \mathcal{V}_2] = \sum_{i,j} \left(a^i \frac{\partial}{\partial x^i} b^j - b^i \frac{\partial}{\partial x^i} a^j \right) \frac{\partial}{\partial x^j} \quad (6.52)$$

This is called the *Lie bracket* of the vector fields.

b.) Let $\text{Der}(U)$ be the space of derivations of $\mathcal{A} = \mathcal{C}^\infty(U)$ valued in $\mathcal{A}$. Show that if we regard $\text{Der}(U)$ as a vector space over $\mathbb{R}$ then the bracket

$$[\cdot, \cdot] : \text{Der}(U) \times \text{Der}(U) \rightarrow \text{Der}(U) \quad (6.53)$$

is bilinear and antisymmetric.

c.) Prove the *Jacobi identity*:

$$[\mathcal{V}_1, [\mathcal{V}_2, \mathcal{V}_3]] + [\mathcal{V}_2, [\mathcal{V}_3, \mathcal{V}_1]] + [\mathcal{V}_3, [\mathcal{V}_1, \mathcal{V}_2]] = 0 \quad (6.54)$$

d.) Show that if we consider $\text{Der}(U)$ as a module over $\mathcal{A}$ then $[\cdot, \cdot]$ is not bilinear. Evaluate $[f_1 \mathcal{V}_1, f_2 \mathcal{V}_2]$ where $f_1, f_2 \in \mathcal{A}$.

Definition: A vector space satisfying (b) and (c) is called a *Lie algebra*.

The historical origin of Lie algebras is indeed associated with vector fields on manifolds.

Exercise *Derivations form a Lie algebra*

Show that if D_1, D_2 are derivations of any algebra $\mathcal{A}$ over k then $[D_1, D_2]$ satisfies the Leibniz rule:

$$[D_1, D_2](fg) = f[D_1, D_2](g) + [D_1, D_2](f)g \quad (6.55)$$

and hence $Der(\mathcal{A})$ is a Lie algebra over κ .

Exercise

Consider the circle as the unit circle in the complex plane. Consider the basis of analytic vector fields $z^{n+1} \frac{d}{dz}$, $n \in \mathbb{Z}$. Compute the Lie brackets of these vector fields.

6.3.3 First definition of T_p^*M

One way to define the cotangent space of M at p , denoted T_p^*M , is as the space of linear functionals on T_pM :

Definition 1 The *cotangent space* to M at p is the vector space:

$$T_p^*M \equiv Hom(T_pM, \mathbb{R}) \quad (6.56)$$

vectors in T_p^*M are called *1-forms at p* .

Given any function $f \in C^1(p)$ there is a tautological cotangent vector

$$df(p) \in T_p^*M \quad (6.57)$$

It is defined by:

$$\langle df(p), V \rangle = df(p)(V) = V(f) \quad \forall V \in T_pM \quad (6.58)$$

In particular, let us choose a coordinate system $\{x^i\}$ in a neighborhood U of p . The x^i are functions and therefore we have differentials $dx^i(p)$.

Theorem: $T_p^*M = \text{Span}_{\mathbb{R}} \{dx^i(p)\}$.

Proof: $\frac{\partial}{\partial x^i} |_p \text{ span } T_pM$. $\langle dx^i(p), \frac{\partial}{\partial x^j} |_p \rangle = \delta_j^i$ so $dx^i(p)$ is the dual basis. ♠

6.3.4 Algebraic definition of T_p^*M

Instead of regarding T_pM as fundamental and T_p^*M as a derived concept we could have worked the other way around, as follows.

Let $\mathfrak{m}_p$ be the subspace of $C^\infty(M; \mathbb{R})$ of functions which *vanish* at p . It is clearly a vector subspace, but it is more, it is what is called a *maximal ideal*. It is an ideal because, for any function $f \in \mathfrak{m}_p$ and any $g \in C^\infty(M; \mathbb{R})$ the product $fg \in \mathfrak{m}_p$. It is maximal, because any vector subspace containing $\mathfrak{m}_p$ must contain a function g with $g(p) \neq 0$. But then we can decompose any function $f \in C^\infty(M; \mathbb{R})$ as $f = \frac{f(p)}{g(p)}g + h$ and $h = f - \frac{f(p)}{g(p)}g \in \mathfrak{m}_p$.

Now, one can usefully define a filtration on the space of all functions:

$$C^\infty(M; \mathbb{R}) \supset \mathfrak{m}_p \supset \mathfrak{m}_p^2 \supset \mathfrak{m}_p^3 \supset \dots \quad (6.59)$$

where $\mathfrak{m}_p^2$ means the subspace of functions which are linear combinations of products of functions in $\mathfrak{m}_p$, and so on. For real analytic functions (i.e. those with a convergent Taylor

series in the neighborhood of p which converges to the value of the function) we can identify $\mathfrak{m}_p^2$ as the space of those functions which vanish together with all their first derivatives at p (in any coordinate system). Similarly, the real analytic functions in $\mathfrak{m}_p^k$ are those with Taylor series (in any coordinate system with $x^i(p) = 0$):

$$f(x) = a_{i_1, \dots, i_k} x^{i_1} \dots x^{i_k} + \dots \quad (6.60)$$

Definition 2: The cotangent space to M at p is $T_p^*M := \mathfrak{m}_p / \mathfrak{m}_p^2$

Now, in these terms we can define the one-form at p associated with a function f to be:

$$df_p := [f - f(p) \cdot 1] \in \mathfrak{m}_p / \mathfrak{m}_p^2 \quad (6.61)$$

Theorem : Suppose $\{x^i\}$ is a coordinate system near p then:

$$\text{Span} \{dx_p^1, \dots, dx_p^n\} = T_p^*M \quad (6.62)$$

Proof: It suffices to consider real analytic functions. So use the Taylor expansion to write (in a coordinate system with $x^i(p) = 0$):

$$f - f(p) \cdot 1 = a_i x^i + a_{ij} x^i x^j + \dots \quad (6.63)$$

therefore

$$\begin{aligned} [f - f(p) \cdot 1] &= [a_i x^i] = a_i [x^i] \\ &= \sum_i a_i dx_p^i. \end{aligned} \quad (6.64)$$

To prove linear independence note that:

$$\begin{aligned} \sum \lambda_i dx_p^i = 0 &\rightarrow \sum \lambda_i x^i \in \mathfrak{m}_p^2 \\ &\rightarrow \frac{\partial}{\partial x^j} \Big|_0 \left(\sum \lambda_i x^i \right) = 0 \\ &\rightarrow \lambda_j = 0 \spadesuit \end{aligned} \quad (6.65)$$

Thus, our second definition of the cotangent space is equivalent to the first definition.

If we consider the cotangent space to be the primary concept then we could define the tangent space as the dual space:

Definition 3: The tangent space T_pM to M at p is the vector space:

$$T_pM := \text{Hom}(T_p^*M, \mathbb{R}) \quad (6.66)$$

Warning: We are implicitly assuming above that the double-dual of a vector space is the original vector space: $(V^*)^* \cong V$. In infinite dimensions, if one requires additional structure such as that V be a Banach space this is not necessarily so.

Exercise

Under change of coordinates

$$dy^i(p) = \frac{\partial y^i}{\partial x^j}(p) dx^j(p) \quad (6.67)$$

Exercise

Let

$$\begin{aligned} \omega_p &= \omega_i dx_p^i \in T_p^*M \\ V_p &= V^i \frac{\partial}{\partial x^i} \big|_p \in T_pM \end{aligned} \quad (6.68)$$

Show

$$\langle \omega_p, V_p \rangle = \sum_{i=1}^n \omega_i V^i \quad (6.69)$$

6.3.5 The differential of a map

Suppose $f : M_1 \rightarrow M_2$ is a differentiable map between two manifolds. There are two canonical linear maps:

1. $f_* : T_p M_1 \rightarrow T_{f(p)} M_2$. This is called the *pushforward map*. It is also called the *differential of f* and is often denoted df_p .²⁷
2. $f^* : T_{f(p)}^* M_2 \rightarrow T_p^* M_1$. This is called the *pullback map*.

We can define f_* simply by noting that any differentiable path λ in M_1 passing through p is mapped to a differentiable path $f \circ \lambda$ in M_2 passing through $f(p)$. We define the differential of f to map the tangent vector to λ to be the tangent vector of this pushed-forward curve:

$$f_* \left(\lambda_* \left(\frac{d}{dt} \right) \right) := (f \circ \lambda)_* \left(\frac{d}{dt} \right) \quad (6.70)$$

In terms of derivations of $\mathcal{A}_i = \mathcal{C}^\infty(M_i)$, we can use the (contravariant) pullback of functions to push forward a derivation D_1 of $\mathcal{A}_1$ to a derivation of $\mathcal{A}_2$ by defining, for all $F \in \mathcal{C}^\infty(M_2)$,

$$D_2(F) := D_1(f^*(F)) = D_1(F \circ f) \quad (6.71)$$

It is easy to check this is a derivation.

²⁷Do not confuse it with the one-form at p defined by a real-valued function!

Dually, if $[F] \in \mathfrak{m}_{f(p)}/\mathfrak{m}_{f(p)}^2 \in T_{f(p)}^*M_2$ then the pullback is defined by pulling back any representative function:

$$f^*[F] := [F \circ f] \in \mathfrak{m}_p/\mathfrak{m}_p^2 \in T_p^*M_1 \quad (6.72)$$

One has to check this is really well-defined!

Note that it follows immediately from the definitions that if $\omega \in T_{f(p)}^*M_2$ and $V \in T_pM_1$ then

$$\langle f^*(\omega), V \rangle = \langle \omega, f_*(V) \rangle \quad (6.73)$$

Alternatively, since the pairing of T^*M and TM is nondegenerate, given a definition of one of f^* or f_* we can define the other from this formula.

Let us express these maps in local coordinates. Choose a coordinate chart (U_1, ϕ_1) near $p \in M_1$ and (V_2, ψ_2) near $f(p)$ in M_2 . Then

$$\psi_2 \circ f \circ \phi_1^{-1} : \phi_1(U) \subset \mathbb{R}^{n_1} \rightarrow \psi_2(f(U) \cap V_2) \subset \mathbb{R}^{n_2} \quad (6.74)$$

will be expressed as an n_2 -tuple of functions

$$y^a(x^1, \dots, x^{n_1}) \quad a = 1, \dots, n_2 \quad (6.75)$$

In these terms then we have

$$f_* \left(\frac{\partial}{\partial x^i} \right) = \sum_{a=1}^{n_2} \frac{\partial y^a}{\partial x^i} \frac{\partial}{\partial y^a} \quad (6.76)$$

$$f^*(dy^a) = \sum_{i=1}^{n_1} \frac{\partial y^a}{\partial x^i} dx^i \quad (6.77)$$

The matrix $\frac{\partial y^a}{\partial x^i}$ is a matrix representation of f^* and f_* in a coordinate basis. If we work with f_* we should regard it as an $n_2 \times n_1$ matrix. If we work with f^* we should regard it as an $n_1 \times n_2$ matrix, related to the previous one by transpose. Under change of coordinate chart it is multiplied on the left or right by the appropriate invertible matrix $J[\phi'_1; \phi_1]$ or $J[\psi'_2; \psi_2]$.

Remark: Since $\frac{d}{dt}$ is a natural tangent vector to a point $t \in (a, b)$ this explains our strange notation $\lambda_* \left(\frac{d}{dt} \right)$ for the tangent to the curve $\lambda(t)$ in M .

6.4 Orientability

Let V be a real vector space and let $GL(V)$ be the of $\mathbb{R}$ -linear invertible transformations. $GL(V)$ is a topological group and there is a continuous map $\det : GL(V) \rightarrow \mathbb{R}^*$. Let $GL^+(V)$ be the subgroup of linear transformations of positive determinant. Note that $GL^+(V)$ acts on the set $\mathcal{B}$ of ordered bases for V .

Definition: An *orientation* on V is an equivalence class under the equivalence relation on $\mathcal{B}$ defined by $b' \sim b$ if there is a $g \in GL^+(V)$ with $b' = g \cdot b$.

Any two bases $b', b \in \mathcal{B}$ are related by *some* $GL(V)$ transformation, so $b' = g \cdot b$ for some $g \in GL(V)$. Accordingly there are just two equivalence classes: Any real vector space has just two possible orientations.

Upon choosing a basis for V we have an isomorphism $GL(V) \cong GL(n; \mathbb{R})$. A basis for $\mathbb{R}^n$ is an ordered set $\{v_1, \dots, v_n\}$ of column vectors. These can be used to form the columns of a matrix and if the determinant of that matrix is positive then the basis is in the equivalence class of the standard basis $\{e_1, \dots, e_n\}$ and if it is negative it is in the other equivalence class.

Remark: In fact $GL(n; \mathbb{R})$ is a union of disjoint open sets: $GL^+(n; \mathbb{R}) \amalg GL^-(n; \mathbb{R})$. and $GL^\pm(n; \mathbb{R})$ are the two connected components.

Now consider the tangent space of an n -dimensional manifold at a point p . Since $T_p M$ is a real vector space it has two possible orientations. Choose an orientation on $T_p M$. Then, a coordinate chart (U, ϕ) in a neighborhood of p has a *compatible orientation* if the isomorphism

$$T_p M \cong \mathbb{R}^n \quad (6.78)$$

provided by the basis $\frac{\partial}{\partial x^1}, \dots, \frac{\partial}{\partial x^n}$ is orientation-preserving. That is, if the coordinate basis $\frac{\partial}{\partial x^1}, \dots, \frac{\partial}{\partial x^n}$ is in the equivalence class of the chosen orientation of $T_p M$. Now, if p and q are both in U then we say that they have compatible orientations if there is a coordinate system which has a compatible orientation with both $T_p M$ and $T_q M$.

Definition: A manifold M is *orientable* if there is a compatible choice of orientations on all the tangent spaces $T_p M$. It is *unorientable* if there is no such choice. An *orientation* on an orientable manifold M is a choice of such a compatible set of orientations. If M is equipped with an orientation it is said to be *oriented*.

If a manifold is orientable then there is a choice of coordinate atlas with transition functions such that

$$\det J[\phi_2; \phi_1] = \det\left(\frac{\partial \phi_2^i}{\partial x^j}\right) > 0$$

on all overlaps.

Remarks:

1. Every manifold M , orientable or not, has a canonically oriented double-cover $\widehat{M}$. We cover each chart U by a $2 : 1$ cover where the extra discrete label is an orientation on $\mathbb{R}^n$. Thus, the oriented double cover of $\mathbb{R}^n$ is $\mathbb{R}^n \amalg \mathbb{R}^n$ where the first summand carries the standard orientation and the second one carries the other orientation. Similarly, the oriented double cover of S^2 is, as a manifold $S^2 \amalg S^2$ where each of the two summands carries one of the two possible orientations on S^2 . Now note that the oriented double cover of $\mathbb{RP}^2$ is a copy of S^2 with an orientation. The oriented double cover of the Klein bottle is a copy of a torus. A useful general result is:

Theorem: Let M be a connected manifold. Then the oriented double cover $\widehat{M}$ is *connected* iff M is unorientable.

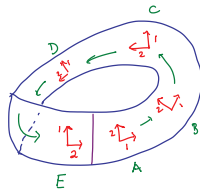


Figure 25: Choose a local orientation and follow it around a closed path on a Möbius strip, as in the path ABCDE. The final orientation will be opposite to the original one. Hence the Möbius strip is unorientable.

2. *Möbius strip.* The canonical example of an unorientable manifold is the Möbius strip shown in Figure 61. Let take the quotient of $\mathbb{R} \times [0, 1]$ by $(x, y) \sim (x + 2, y)$. This gives a cylinder. Take a further quotient by $(x, y) \sim (x + 1, 1 - y)$. This gives the Möbius strip. Thus the cylinder is a double-cover of the Möbius strip and is the oriented double cover. See Figure 25.

6.5 Tangent and cotangent bundles

The tangent space $T_p M$ assigns a vector space to each point p on M . It is clear, in an intuitive sense, that the vector spaces vary continuously with p . Given a coordinate patch (U, ϕ) around p we can identify the union of tangent spaces $T_q M$ for $q \in U$ with $U \times \mathbb{R}^n$, where $n = \dim M$. This notion can be globalized to define a space called the *tangent bundle* of M , together with the associated *cotangent bundle*. As sets these are defined by:

$$\begin{aligned} T^*M &= \cup_{p \in M} T_p^*M \\ TM &= \cup_{p \in M} T_p M \end{aligned} \tag{6.79}$$

The spaces T^*M and TM can be regarded as families of vector spaces parametrized by $p \in M$. They have some very interesting mathematical structures associated with them.

First of all, TM and T^*M are manifolds. Choose an atlas $\{(U_\alpha, x_\alpha^\mu)\}$ for M . Then, thanks to the coordinates we have a homeomorphism

$$\cup_{q \in U_\alpha} T_q M \cong U_\alpha \times \mathbb{R}^n \tag{6.80}$$

Indeed, in the patch U_α we have bases:

$$\begin{aligned} \left. \frac{\partial}{\partial x_\alpha^\mu} \right|_p & \text{ for } T_p M \quad \forall p \in U_\alpha \\ dx_\alpha^\mu(p) & \text{ for } T_p^* M \quad \forall p \in U_\alpha \end{aligned} \tag{6.81}$$

So a cotangent vector

$$\omega = \omega_\mu dx^\mu \quad (6.82)$$

is specified by coordinates

$$\{(x^1, \dots, x^n; \omega_1, \dots, \omega_n)\} \quad (6.83)$$

Note: At this point we have made a conceptual jump: ω_μ was formerly regarded as a component of the expansion of a 1-form ω with respect to the basis dx^μ . It is now regarded as a coordinate on another manifold, T^*M . Note that $\dim T^*M = 2\dim M$.

So: The patches for T^*M are

$$U_\alpha \times \mathbb{R}^n \quad \text{coordinates} \quad \{(x^\mu_\alpha, \omega^\alpha_\mu)_{\mu=1\dots n}\} \quad (6.84)$$

transition functions: we patch together $U_\alpha \times \mathbb{R}^n$ with $U_\beta \times \mathbb{R}^n$ by

$$\omega^\alpha_\mu dx^\mu_\alpha = \omega^\beta_\mu dx^\mu_\beta \quad (6.85)$$

therefore, although the transition functions x^μ_α are nonlinear functions of the x^ν_β on patch overlaps the transition functions for the coordinates ω_μ is

$$\omega^\beta_\mu = \sum_{\nu=1}^n \frac{\partial x^\nu_\alpha}{\partial x^\mu_\beta} \omega^\alpha_\nu \quad (6.86)$$

Note that this is a linear transformation of the ω_μ coordinates. This is the extra structure which allows us to do linear algebra in families.

The space T^*M has a canonical projection map $\pi : T^*M \rightarrow M$. If $\omega \in T^*_p M$ then $\pi(\omega) = p$. Note that the fiber of the map $\pi^{-1}(p)$ is a vector space. On patch overlaps the transition functions are linear transformations on the fibers: These are the defining properties of a *vector bundle*. See Section **** below for a more formal description of bundles.

- The Jacobian matrix across patch boundaries has the form

$$J(x^\mu_\beta, \omega^\beta_\mu; x^\nu_\alpha, \omega^\alpha_\nu) = \begin{bmatrix} \frac{\partial x^\mu_\beta}{\partial x^\nu_\alpha} & 0 \\ * & \frac{\partial x^\nu_\alpha}{\partial x^\mu_\beta} \end{bmatrix} \quad (6.87)$$

Examples

1. $T^*\mathbb{R}^n = \mathbb{R}^n \times \mathbb{R}^n = \mathbb{R}^{2n} = \{(x^\mu, p_\mu)\}$. There is a similar description of $T\mathbb{R}^n$.
2. $T^*S^1 = \text{cylinder}$. Coordinates (θ, p) :

Remark In describing the classical mechanics of a particle moving on a manifold M the phase space is T^*M . This is very useful both in classical mechanics and also in the transition to the quantum mechanics of a particle moving on M .

Figure 26: The cylinder can be regarded as the cotangent bundle T^*S^1 .

Exercise

Describe TM in an analogous way.

Exercise

Show that TM and T^*M are noncompact manifolds which are contractible to M .

Exercise

a.) If $v \in T_pM$ show that there are natural isomorphisms:

$$T_v(TM) \cong T_pM \oplus T_pM \quad (6.88)$$

b.) If $\omega \in T_p^*M$ show that there are natural isomorphisms:

$$T_\omega(T^*M) \cong T_pM \oplus T_p^*M \quad (6.89)$$

Exercise *Canonical trivialization of the Liouville form*

Show that there is a globally defined 1-form on T^*M .

Choose an atlas $\{(\mathcal{U}_\alpha, \phi_\alpha)\}$ for M and let $x_\alpha^\mu := x^\mu(\phi_\alpha)$ be the coordinate function on $\mathcal{U}_\alpha$. Define a collection of functions p_μ^α on $\pi^{-1}(\mathcal{U}_\alpha) \subset T^*M$ by declaring

$$p_\mu^\alpha(\omega) := \omega_\mu^\alpha \quad (6.90)$$

if $\omega = \omega_\mu^\alpha dx_\alpha^\mu$.

Show that $\lambda := p_\mu^\alpha(\omega) dx_\alpha^\mu$ is a globally well-defined one-form on T^*M .

Remark: This one-form has the important property that it trivializes the canonical symplectic form on T^*M , that is $\omega = d\lambda$.

Exercise

Consider the affine algebraic variety:

$$\{\vec{z} : z_0^2 + z_1^2 + \cdots + z_n^2 = \mu^2\} \subset \mathbb{C}^{n+1} \quad (6.91)$$

Show that if $\mu \neq 0$ this space is diffeomorphic to T^*S^{2n+1} .

Exercise *Orientability*

Show that T^*M and TM are always orientable manifolds.

6.5.1 Example 1: TS^2 and T^*S^2

We can easily work out the transition functions. Let us start with T^*S^2 , identify $S^2 \cong \mathbb{CP}^1$ and work in complex coordinates, identifying the fiber $\mathbb{R}^2$ with $\mathbb{C}$ as a rank 2 real vector space. Locally a cotangent vector can be written as

$$\omega = p_{\pm} dz_{\pm} + p_{\pm}^* dz_{\pm}^* \quad (6.92)$$

Then $z_- = 1/z_+$ and the overlap conditions must be

$$p_+ dz_+ = p_- dz_- \quad (6.93)$$

so

$$p_+ = -1/(z_+)^2 p_- \quad (6.94)$$

Translating back to real coordinates, $p_{\pm} = \omega_{\pm}^{\pm} + i\omega_{\pm}^{\mp}$. we have $(x_+; (\omega_+^1, \omega_+^2))$ glued to $(x_+; (\omega_+^1, \omega_+^2))$ by

$$\begin{pmatrix} \omega_+^1 \\ \omega_+^2 \end{pmatrix} = - \begin{pmatrix} \cos 2\phi & -\sin 2\phi \\ \sin 2\phi & \cos 2\phi \end{pmatrix} \begin{pmatrix} \omega_-^1 \\ \omega_-^2 \end{pmatrix} \quad (6.95)$$

FIGURE

The dual space is obtained by taking $\phi \rightarrow -\phi$ above.

Alternatively, we could use coordinates $z_{\pm} = x_{\pm}^1 + ix_{\pm}^2$ and compute:

$$\frac{1}{2}(\partial_1^+ - i\partial_2^+) = \frac{\partial}{\partial z_+} = \frac{\partial z_-}{\partial z_+} \frac{\partial}{\partial z_-} \quad (6.96)$$

to get the transition functions on the equator:

$$\begin{pmatrix} \partial_1^+ \\ \partial_2^+ \end{pmatrix} = - \begin{pmatrix} \cos 2\phi & \sin 2\phi \\ \sin 2\phi & \cos 2\phi \end{pmatrix} \begin{pmatrix} \partial_1^- \\ \partial_2^- \end{pmatrix} \quad (6.97)$$

Remarks:

1. From the above discussion it is clear that TS^2 can also be regarded as a holomorphic line bundle over $\mathbb{C}P^1$. In fact, it is isomorphic to $\mathcal{L}^2$ where $\mathcal{L}$ is the tautological bundle. Thus T^*S^2 is isomorphic to $\mathcal{L}^{-2}$.
2. As a real noncompact 4-manifold T^*S^2 is a very interesting manifold that admits a nontrivial solution of the self-dual Einstein equations. It is the first example of an ALE space.
3. The fact that TS^2 is a nontrivial bundle shows that you can't comb the hair on a sphere.

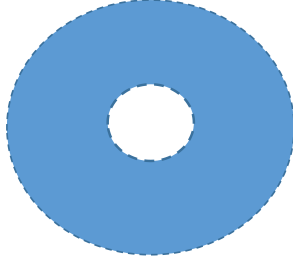


Figure 27: An embedding of the cylinder into the plane.

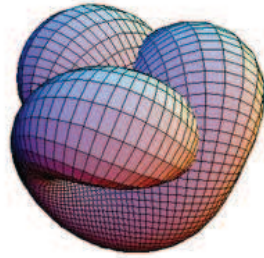


Figure 28: Boy's surface, an immersion of $\mathbb{RP}^2$ into $\mathbb{R}^3$. From Wikipedia.

6.6 Definitions: Submersion, immersion, embedding, critical and regular points and values

Definition: Let $f : M_1 \rightarrow M_2$ be a differentiable map of manifolds. Then

1. f is a *submersion at a point p* if the linear map $df : T_p M_1 \rightarrow T_{f(p)} M_2$ is surjective. If f is a submersion at all points $p \in M_1$ it is simply said to be a *submersion*.

2. If f is a submersion at $p \in M_1$ then the point is called a *regular point of f* .
3. If f is not a submersion at $p \in M_1$, that is, if df_p has rank less than $n_2 = \dim M_2$ then p is said to be a *singular point of f* or, equivalently, a *critical point of f* .²⁸ If q is in the image of f and one of its preimages $p \in f^{-1}(q)$ is a critical point, then q is said to be a *critical value of f* . If q is not a critical value of f it is said to be a *regular value of f* .²⁹
4. f is an *immersion at a point p* if the linear map $df : T_p M_1 \rightarrow T_{f(p)} M_2$ is injective, i.e. has no kernel. If f is an immersion at all points $p \in M_1$ it is simply said to be an *immersion*.
5. f is an *embedding* if it is an immersion which is also injective, and moreover a topological embedding, i.e. a homeomorphism onto its image.
6. A *submanifold* of a manifold M is a manifold Σ which is also a subspace $\Sigma \subset M$ with the subspace topology such that $\Sigma = f(N)$ is the image of an embedding of another manifold $f : N \rightarrow M$.

Examples

1. Consider maps of S^1 into $\mathbb{R}^2$. The standard unit circle is an embedding. A figure 8 is an immersion, but not an embedding. In both these cases if we choose a standard parametrization of S^1 , say $z = e^{2\pi i t}$ then the tangent vector $\frac{d}{dt}$ is mapped to a tangent vector $\lambda_*\left(\frac{d}{dt}\right)$ to the curve. The condition that λ is an immersion is the condition that the velocity is always nonzero.
2. A *knot* is an embedding of S^1 into $\mathbb{R}^3$. A *singular knot* is an immersion of S^1 into $\mathbb{R}^3$. Thus, the classification of knots is a classification of embeddings of S^1 into $\mathbb{R}^3$ up to isotopy. The space of all immersions is carved up into chambers.
3. The cylinder can be embedded in $\mathbb{R}^2$. See Figure 27. However the Mobius strip cannot be embedded into the plane! Note that in Figure 25 there are self-intersections! This reflects the fact that the tangent bundle of the cylinder is trivial while that of the Mobius strip is not.
4. A famous immersion of $\mathbb{RP}^2$ into $\mathbb{R}^3$ is known as Boy's surface. Figure 28.³⁰ An immersion of the Klein bottle into $\mathbb{R}^3$ is shown in Figure 65. One can prove that $\mathbb{RP}^2$

²⁸Some authors, notably V. Arnold et. al. define a critical point to be a point where df_p has a rank $r < \min[n_1, n_2]$. This is not often used, but Arnold's work on critical points and singularity theory has been very influential.

²⁹Note that, in strict logic, this means that any $q \in M_2$ which is not in the image of f is considered to be a "regular value of f " even though it is not a value of f .

³⁰For an extended discussion see Hilbert and Cohn-Vossen, *Geometry and the Imagination*, pp. 317-321. Hilbert's student Werner Boy found the surface in response to a problem posed by Hilbert. Hilbert asked him to show that the projective plane could not be immersed in $\mathbb{RP}^2$! There is a sense ("immersion cobordism") in which 8 copies of Boy's surface can be shrunk to a point. See the wonderful video "Eight Boy's Bound" at <https://www.youtube.com/watch?v=7ZbbhBQEJmI>.

and the Klein bottle, cannot be embedded into $\mathbb{R}^3$. The proof relies on the topology of the tangent bundle to these surfaces.

Exercise *The Hessian*

Let $f : M \rightarrow \mathbb{R}$. Consider the matrix of second derivatives

$$\frac{\partial^2 \tilde{f}}{\partial x^i \partial x^j} \quad (6.98)$$

where $\tilde{f} = f \circ \phi^{-1}$.

- a.) Compute how the matrix changes under a change of coordinates.
 - b.) How does the formula simplify when x is a critical point of f ? The matrix of second derivatives at a critical point is known as the *Hessian* of f .
-

Exercise

- a.) Show that the map $f : \mathbb{RP}^2 \rightarrow \mathbb{R}^3$ defined by

$$f([X^1 : X^2 : X^3]) = \frac{(X^2 X^3, X^1 X^3, X^1 X^2)}{(X^1)^2 + (X^2)^2 + (X^3)^2} \quad (6.99)$$

is well-defined and smooth.

- b.) Where does it fail to be an immersion?
-

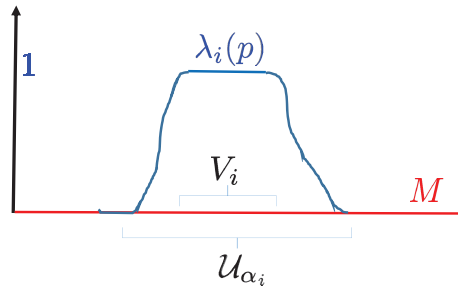


Figure 29: A collection of functions on M used to prove the Whitney embedding theorem.

6.7 Whitney embedding theorem

Whitney embedding theorem. Every compact manifold of dimension n can be embedded into Euclidean space $\mathbb{R}^{2n}$.

Partial proof: ³¹ The proof proceeds in three steps, and we will only indicate some of it, to get some of the key ideas.

Step 1: Prove that any compact manifold of dimension n can be embedded into $\mathbb{R}^N$ for some (sufficiently large) N .

Step 2: Use a reduction procedure to show that if a manifold $M \hookrightarrow \mathbb{R}^N$ and $N > 2n+1$ then we can construct another embedding into $\mathbb{R}^{N-1}$.

Step 3: Reduce from an embedding into $\mathbb{R}^{2n+1}$ to $\mathbb{R}^{2n}$. This last step is quite hard and we will skip it entirely.

Proof A of step 1: Proof A assumes some knowledge of vector bundles. If $E \rightarrow M$ is a vector bundle over a compact manifold M then there is a complementary bundle $E^\perp \rightarrow M$ such that

$$E \oplus E^\perp \cong M \times \mathbb{R}^N \quad (6.100)$$

for some sufficiently large N . Let us apply this to $E = TM$. Then

$$TM \oplus E^\perp \cong M \times \mathbb{R}^N \quad (6.101)$$

There certainly exist sections $s : M \rightarrow TM$ so that $s(TM)$ is transverse to the zero section, so then $s(TM)$ is an embedded submanifold in the total space of TM . Now, via the isomorphism (6.101), the embedding of M into TM can be considered to be an embedding of M into the fixed fiber $\mathbb{R}^N$ for some sufficiently large N .

Proof B of step 1: Using partitions of unity one can show that M has a finite atlas $\{(U_\alpha, \phi_\alpha)\}$ and a finite subcover $\{V_i\}_{i=1}^K$ so that for all i there is an α_i with $\bar{V}_i \subset U_{\alpha_i}$, and, moreover, there is a collection of functions $\lambda_i : M \rightarrow \mathbb{R}$ with $\lambda_i(p) = 1$ for $p \in V_i$ and while the support of λ_i is contained in the corresponding U_{α_i} . See Figure 29.

Now, if $p \in V_i$ define functions $\psi_i : M \rightarrow \mathbb{R}$ by $\psi_i(p) := \lambda_i(p)\phi_{\alpha_i}(p)$. Note that a clever maneuver has been made here! The coordinates ϕ_{α_i} are only defined on the local patch U_{α_i} . However, thanks to the function λ_i they can be extended globally as the zero function. This can even be done as a $\mathcal{C}^\infty$ function. Now, the embedding will be:

$$F : p \mapsto (\psi_1(p), \dots, \psi_K(p), \lambda_1(p), \dots, \lambda_K(p)) \in \mathbb{R}^{nK+K} \quad (6.102)$$

We show it is an embedding by first showing it is an immersion and then showing it is 1-1.

To see it is an immersion

$$dF = d\psi_1 \oplus \dots \oplus d\psi_K \oplus d\lambda_1 \oplus \dots \oplus d\lambda_K \quad (6.103)$$

Every point p is in V_i for some i and then $d\psi_i = d\phi_{\alpha_i}$ has zero kernel since ϕ_{α_i} is a system of coordinates. To see it is one-one, suppose $F(p) = F(q)$. Then $\lambda_j(p) = \lambda_j(q)$ for all j .

³¹We follow Bredon's book here.

Now for some i , we must have $p \in V_i$. But then $\psi_i(p) = \psi_i(q)$ implies $\phi_i(p) = \phi_i(q)$ since ϕ_i is a homeomorphism on U_{α_i} we must have $p = q$.

Remark: In fact, a similar argument is used to prove (6.100).

Proof B of step 2: We can now consider $M \subset \mathbb{R}^N$. Suppose there is a vector w in $\mathbb{R}^N$ such that

(*) If $x, y \in M$ with $x \neq y$ then w is not parallel to $x - y$.

Now we take $M \cap w^\perp$. The projection $x \rightarrow x - (x, w)w/(w, w)$ is an embedding under condition (*). In this way, given such a w we can reduce N by 1.

Now consider the map:

$$\tau : M \times M - \{(x, x) : x \in M\} \rightarrow \mathbb{R}P^{N-1} \quad (6.104)$$

defined via $(x, y) \rightarrow [x - y]$ then a vector $[w]$ not in the image of τ satisfies criterion (*).

Since τ is defined on a manifold of dimension $2\dim M = 2n$, as long as $2n < N - 1$ we can always find such a w .³² The process stops when $2n = N - 1$, ie $N = 2n + 1$. ♠.

Remarks

1. There is also a Whitney immersion theorem, which states that any compact n -dimensional manifold can be immersed in Euclidean space of dimension $2n - 1$.
2. The values $2n$ for embedding and $2n - 1$ for immersion are the best possible. One proves this by studying the topology of the normal bundle of an embedding and its topological invariants, known as *Stiefel-Whitney classes*. See, for example Milnor and Stasheff, *Characteristic Classes* for a clear exposition of this application.

Exercise

Show that the boundary is a submanifold of dimension $n - 1$.

6.8 Local form of maps between manifolds and of submanifolds

6.8.1 Local form of a differentiable map between manifolds

Suppose that $f : M_1 \rightarrow M_2$ is a differentiable map between two manifolds. By choosing suitable coordinates (U, ϕ) around any point $p \in M_1$ and (V, ψ) around its image $f(p) \in M_2$ one can put the local form of the function $\psi \circ f \circ \phi^{-1}$ in a very simple form. This is quite often useful for doing local computations.

The key theorem is based on an important piece of advanced calculus:

Theorem[Inverse Function Theorem]. Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a $\mathcal{C}^1$ -function in an open set containing p and suppose that

$$\det df_p = \det \frac{\partial f^a}{\partial x^i}(p) \neq 0 \quad (6.105)$$

³²This can be made precise using “Sard’s theorem.” For more details see G. Bredon, pp. 89-92.

Then there is an open set U around p and an open set V around $f(p)$ so that $f : U \rightarrow V$ has a continuously differentiable inverse $f^{-1} : V \rightarrow U$.

For a proof see M. Spivak, *Calculus on Manifolds*, Theorem 2-11,

Now suppose that $f : M_1 \rightarrow M_2$ is differentiable in a neighborhood U of p . We will aim to simplify the coordinate expression for f as much as possible using the freedom to make differentiable change of coordinates in some suitable neighborhoods of p and $f(p)$.

To begin, choose arbitrary coordinate charts $(\tilde{U}, \tilde{\phi})$ around p and $(\tilde{V}, \tilde{\psi})$ around $f(p)$. Then, in these coordinates we can write:

$$\tilde{f} = \tilde{\psi} \circ f \circ \tilde{\phi}^{-1} : (u^1, \dots, u^{n_1}) \mapsto (v^1, \dots, v^{n_2}) \quad (6.106)$$

where $v^a = v^a(u^1, \dots, u^{n_1})$ is an n_2 -tuple of functions of n_1 variables. WLOG we can and will assume that $u^i(p) = 0$ and $v^a(f(p)) = 0$. When we write u, v, x, y without indices it means the whole tuple.

Now, let us suppose that $df_p : T_p M_1 \rightarrow T_{f(p)} M_2$ has rank r . Linear algebra demands that we must have $r \leq \min[n_1, n_2]$. Then, by suitably reordering the u^i and v^a we can guarantee that the $r \times r$ matrix:

$$\left. \frac{\partial v^\alpha}{\partial u^\beta} \right|_0 \quad 1 \leq \alpha, \beta \leq r \quad (6.107)$$

is invertible. Now consider the map $\Omega_1 : U' \subset \mathbb{R}^{n_1} \rightarrow U'' \subset \mathbb{R}^{n_1}$

$$\Omega_1 : (u^1, \dots, u^{n_1}) \rightarrow (v^1, \dots, v^r, u^{r+1}, \dots, u^{n_1}) \quad (6.108)$$

Here U', U'' are some neighborhoods of $0 \in \mathbb{R}^{n_1}$. The Jacobian of Ω_1 is of the form:

$$\frac{\partial \Omega_1^j}{\partial u^i} = \begin{pmatrix} \frac{\partial v^\alpha}{\partial u^\beta} & 0 \\ * & 1 \end{pmatrix} \quad (6.109)$$

and therefore has nonzero determinant at $u = 0$. We choose U', U'' in accord with the inverse function theorem so we can speak of Ω_1^{-1} . Therefore, by the inverse function theorem we have for $x \in U$, where U is some suitable neighborhood of 0 inside U''

$$F(x) := \tilde{f} \circ \Omega_1^{-1} : (x^1, \dots, x^r, x^{r+1}, \dots, x^{n_1}) \rightarrow (x^1, \dots, x^r, \psi^{r+1}(x), \dots, \psi^{n_2}(x)) \quad (6.110)$$

where we define the functions $\psi^k(x) = v^k(\Omega_1^{-1}(x))$ for $k = r+1, \dots, n_2$. The reason is that

$$(y^1, \dots, y^{n_1}) = \Omega_1^{-1}(x^1, \dots, x^r, x^{r+1}, \dots, x^{n_1}) \quad (6.111)$$

iff

$$(v^1(y), \dots, v^r(y), y^{r+1}, \dots, y^{n_1}) = (x^1, \dots, x^r, x^{r+1}, \dots, x^{n_1}) \quad (6.112)$$

and hence $v^\alpha(y) = x^\alpha$, for $1 \leq \alpha \leq r$. Therefore

$$\begin{aligned}
F(x) &= \tilde{f} \circ \Omega_1^{-1}(x^1, \dots, x^r, x^{r+1}, \dots, x^{n_1}) \\
&= \tilde{f}(y^1, \dots, y^{n_1}) \\
&= (v^1(y), \dots, v^r(y), v^{r+1}(y), \dots, v^{n_2}(y)) \\
&= (x^1, \dots, x^r, \psi^{r+1}(x), \dots, \psi^{n_2}(x))
\end{aligned} \tag{6.113}$$

Equation (6.110) is already of interest by itself. Note that we did not need to change coordinates on the target manifold. Just by a change of coordinates on the domain manifold we have simplified the map in an interesting way so that its coordinate expression is $F(x)$ as above.

Now let us add the further assumption that df_q has rank r for all q throughout some open neighborhood $q \in U$ of p . Then we can simplify the coordinate expression for f much more. Note that

$$\frac{\partial F^a}{\partial x^i} = \begin{pmatrix} 1 & 0 \\ * & \frac{\partial \psi^k}{\partial x^j} \end{pmatrix} \tag{6.114}$$

Here $1 \leq a \leq n_2$ is a row index and $1 \leq i \leq n_1$ is a column index, so in the lower right block $r+1 \leq k \leq n_2$ and $r+1 \leq j \leq n_1$. Now, if df_q has rank r throughout some neighborhood then $dF(x)$ has rank r throughout the corresponding neighborhood of 0 and hence in this neighborhood $\frac{\partial \psi^k}{\partial x^j} = 0$ for $r+1 \leq j \leq n_1$. That is, ψ^k , which is *a priori* a function of n_1 variables, is in fact only a function of the first r variables $x^1, \dots, x^r$.

Now define a second differentially invertible map $\Omega_2 : V' \subset \mathbb{R}^{n_2} \rightarrow V'' \subset \mathbb{R}^{n_2}$ by

$$\Omega_2(y^1, \dots, y^{n_2}) \rightarrow (y^1, \dots, y^r, y^{r+1} - \psi^{r+1}(y^1, \dots, y^r), \dots, y^{n_2} - \psi^{n_2}(y^1, \dots, y^r)) \tag{6.115}$$

This makes sense, because $r \leq n_2$.

Then

$$\Omega_2 \circ F = \Omega_2 \circ \tilde{f} \circ \Omega_1^{-1} : (x^1, \dots, x^r, x^{r+1}, \dots, x^{n_1}) \rightarrow (x^1, \dots, x^r, 0, \dots, 0) \tag{6.116}$$

Summarizing, we have proven ³³

Theorem [Constant rank theorem] If $f : M_1 \rightarrow M_2$ has a differential of constant rank r in some neighborhood of $p \in M_1$ then there exist coordinate charts (U, ϕ) around $p \in M_1$ and (V, ψ) around its image $f(p) \in M_2$ so that in these coordinates f has the form

$$f : (x^1, \dots, x^{n_1}) \mapsto \begin{cases} (0, \dots, 0) & r = 0 \\ (x^1, \dots, x^r, 0, \dots, 0) & 0 < r < n_2 \\ (x^1, \dots, x^{n_2}) & r = n_2 \end{cases} \tag{6.117}$$

Remarks

³³We have loosely followed the discussion in M. Spivak, *Introduction to Differential Geometry*, vol. 1, Theorem

1. In the first case of (6.117) f is locally constant. In the third case we necessarily have $n_2 \leq n_1$.
2. An important and immediate corollary of this theorem is that if f is an immersion then $r = n_1$, so $n_1 \leq n_2$, so we have case 1 if $n_1 = 0$ (i.e. the embedding of a single point), case 2 if $0 < n_1 < n_2$, and case three if $n_1 = n_2$.
3. As we will see later, the equation (6.116), in the $\mathcal{C}^\infty$ context, can be interpreted as a statement about the orbits of the action of the group $\text{Diff}(M_1) \times \text{Diff}(M_2)$ on the space of differentiable maps from M_1 to M_2 .

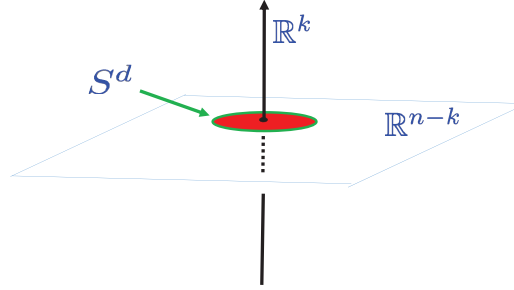


Figure 30: Locally, a k -dimensional submanifold of an n -dimensional manifold is simply determined as the zero-locus of the set of $(n - k)$ coordinates. Therefore, there is a transverse space $\cong \mathbb{R}^{n-k}$ and the unit sphere (in some metric) in that transverse space is diffeomorphic to $S^d = S^{n-k-1}$. Such a sphere is called a linking sphere.

6.8.2 Local form of a submanifold

We can now apply the constant rank theorem to get a very useful local picture of any submanifold $\Sigma \subset M$. Suppose $\dim M = n$ and $\dim \Sigma = k$, then since a submanifold is, by definition, the image of an embedded manifold, and in particular is immersed, we can say that:

Locally, any submanifold can be described - in suitable coordinates - by simply setting some coordinate functions to zero.

That is, locally around any $p \in \Sigma$ there is a coordinate system (U, ϕ) for M such that Σ is the set

$$\phi(U \cap \Sigma) = \{x \in U' \mid x^{k+1} = \dots = x^n = 0\} \quad (6.118)$$

See Figure 30. We refer to the very important quantity

$$\text{cod}(\Sigma) := \dim M - \dim \Sigma \quad (6.119)$$

as the *codimension* of the submanifold. It is often a more useful quantity than the dimension of Σ itself, as we will soon see.³⁴

Figure 30 suggests that given a submanifold $\Sigma \subset M$ there is a single open set $N(\Sigma)$ of M containing Σ which locally looks like $\mathbb{R}^k \times \mathbb{R}^{n-k}$ with the submanifold being $\mathbb{R}^k \times \{0\}$. This is proved in books on differential topology and such a neighborhood of Σ is called a *tubular neighborhood*.

Exercise

Suppose $f : M_1 \rightarrow M_2$ is a $\mathcal{C}^\infty$ function. Show that the restriction of f to a submanifold P of M_1 is also a $\mathcal{C}^\infty$ function.

6.8.3 Defining submanifolds by equations

We can also use the constant rank theorem to give a useful local form of a submersion. In this case df_p has rank n_2 and hence we necessarily have $n_1 \geq n_2$. Then, if f is a submersion in an open set around a point p there are coordinates so that it has the form:

$$f : (x^1, \dots, x^{n_1}) \mapsto (x^1, \dots, x^{n_2}) \quad (6.120)$$

We can, WLOG, assume our coordinate charts are such that f is literally the map $f : \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{n_2}$ keeping only the first n_2 coordinates. Then, every point in the target $\mathbb{R}^{n_2}$ is a regular value of f and the inverse image of a regular value is

$$f^{-1}(c^1, \dots, c^{n_2}) = \{x \in \mathbb{R}^{n_1} | x^1 = c^1, \dots, x^{n_2} = c^{n_2}\} \cong \mathbb{R}^{n_1 - n_2} \quad (6.121)$$

is a submanifold of dimension $n_1 - n_2$.

Therefore we have

Theorem: [Preimage Theorem] If $f : M_1 \rightarrow M_2$ and $q \in M_2$ is a regular value in the image of f then the preimage $f^{-1}(q)$ is a submanifold of M_1 of dimension $\dim M_1 - \dim M_2$. That is, the preimage $f^{-1}(q)$ is a submanifold of M_1 of codimension $\dim M_2$. The tangent space to $f^{-1}(q)$ at any point p is $\ker df_p$.

Now we can apply this idea to describe subsets defined by equations. If $f : M \rightarrow \mathbb{R}^\ell$ and $\dim M \geq \ell$ then for $\vec{c} \in \mathbb{R}^\ell$ the sets

$$M_{\vec{c}} := f^{-1}(\vec{c}) = \cap_{i=1}^{\ell} \{p \in M | f^i(p) = c^i\} \quad (6.122)$$

are called level sets. If $\vec{c}$ is a regular value then the level set is a submanifold of M of codimension ℓ . Note that for each i , $f^i : M \rightarrow \mathbb{R}$ so $df^i : T_p M \rightarrow T_{c^i} \mathbb{R} \cong \mathbb{R}$, and hence df^i is a linear functional on $T_p M$. The regularity condition is the condition that these linear functionals are all linearly independent. We say that the functions f^i are *independent*.

³⁴Also, in infinite dimensions it is not unusual to work with finite-codimension submanifolds. Thus both the manifold and the submanifold are of infinite dimension.

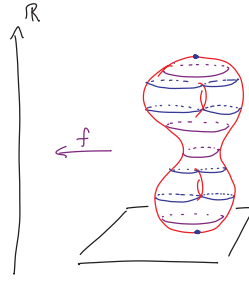


Figure 31: The height function on a Riemann surface. Note that the preimages are submanifolds except at the critical values associated with the two handles.

Example: As an example we consider a single real function. Then the preimages are submanifolds so long as $df_p \neq 0$. A nice example of this is the height function for a Riemann surface sitting on a table as in Figure 31. In this example the Hessian of f at the critical points is nondegenerate. A real-valued function, all of whose critical points have nondegenerate Hessian is called a *Morse function*. In the subject of Morse theory one learns much about the topology of a manifold by studying how the smooth level sets change as the value scans across a critical value. There are many deep connections of physics to Morse theory in the context of supersymmetry.³⁵

All this is worth putting in a slightly different way: Suppose

$$f = (f^1, \dots, f^\ell) : \mathbb{R}^n \rightarrow \mathbb{R}^\ell \quad \ell \leq n \quad (6.123)$$

and p_0 is a regular point of f . By rearranging coordinates, if necessary, we can choose a coordinate system in a neighborhood of p_0 to be of the form

$$(x; y) = (x^1, \dots, x^\ell; y^1, \dots, y^d) \quad (6.124)$$

$\ell + d = n$ so that

$$\frac{\partial f^i}{\partial x^j} \quad 1 \leq i, j \leq \ell \quad (6.125)$$

has rank ℓ at p_0 . Now, if (6.125) has rank ℓ at p_0 it is invertible there, so by the inverse function theorem, for fixed y the map: $F_y : \mathbb{R}^\ell \rightarrow \mathbb{R}^\ell$:

$$F_y : (x^1, \dots, x^\ell) \mapsto (f^1(x, y), \dots, f^\ell(x, y)) \quad (6.126)$$

³⁵See E. Witten, “Supersymmetry and Morse theory,” J. Diff. Geom. **17**, 661 (1982) and K. Hori, S. Katz, A. Klemm, R. Pandharipande, R. Thomas, C. Vafa, R. Vakil and E. Zaslow, “Mirror symmetry,” (Clay mathematics monographs. 1).

is *invertible* in some neighborhood of p_0 with differentiable inverse. Thus, by a change of coordinates we could equally well use the coordinate system:

$$(f^1, \dots, f^\ell, y^1, \dots, y^d) \quad (6.127)$$

in the neighborhood of the level set. Obviously, setting $f^i = c^i$ defines the level set, as predicted by the general local form of a submanifold, as described before.

Taking this a tiny step further gives the

Implicit Function Theorem:

Suppose

$$f : \mathbb{R}^\ell \times \mathbb{R}^d \rightarrow \mathbb{R}^\ell \quad (6.128)$$

has $f(0; 0) = 0$ and

$$\frac{\partial f^i}{\partial x^j} \Big|_{1 \leq i, j \leq \ell} \quad (6.129)$$

has nonzero determinant at $(0; 0)$. Then in an open set $A \times B$ for each $y \in B$ there is a unique $h(y)$ such that

$$f(h(y), y) = 0 \quad (6.130)$$

That is, we can solve for x in terms of y on the level set.

We can prove this using the same ideas used in proving the constant rank theorem. Let the inverse of the map (6.126) be denoted as $x \mapsto g(x; y)$. Note the parametric dependence of g on y . As a map from $\mathbb{R}^\ell \rightarrow \mathbb{R}^\ell$ it is continuously differentiable with continuously differentiable inverse, however, the map $(x, y) \rightarrow (g(x; y), y)$ is also differentiable and hence $g(x; y)$ is also continuously differentiable with respect to y . Finally, by definition:

$$g(f^1(x, y), \dots, f^\ell(x, y); y) = (x^1, \dots, x^\ell) \quad (6.131)$$

and hence the level set $f^i = 0$, for $1 \leq i \leq \ell$ is the same as the set

$$x^i(y) = g^i(0; y) \quad (6.132)$$

That is, we can solve for x in terms of y .

Figure 32: Illustrating the implicit function theorem with the circle.

Example: Consider

$$f(x, y) = x^2 + y^2 - 1 \quad (6.133)$$

then

$$df = (2x, 2y) \quad (6.134)$$

Now consider the circle as the level set $f^{-1}(0)$. In the neighborhood of any point on the level set df is never zero, so the level set is a manifold. Moreover, when $\partial_x f \neq 0$, that is, when $x \neq 0$, that is, when $y \neq \pm 1$, we can use local coordinates (f, y) . Similarly, when $\partial_y f \neq 0$, that is when $x \neq \pm 1$ we can use local coordinates (x, f) . The implicit function theorem guarantees that when $(x, y) \in f^{-1}(0)$, if $\partial_x f \neq 0$ then we can solve for x in terms of y on the level set:

$$x = \pm \sqrt{1 - y^2} \quad (6.135)$$

and this is single valued and differentiable. (The sign depends on which point we use.) Of course this fails precisely in a neighborhood of $y = \pm 1$. Similarly, if $\partial_y f \neq 0$ then we can solve for y in terms of x :

$$y = \pm \sqrt{1 - x^2} \quad (6.136)$$

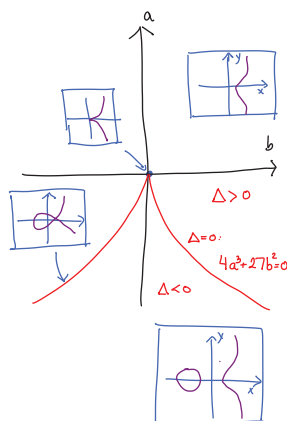


Figure 33: The behaviors of the real cubic curve, $y^2 - (x^3 + ax + b) = 0$ as a function of the parameters a, b

Example 2 Consider $F : \mathbb{R}^2 \rightarrow \mathbb{R}$ with

$$F(x) = y^2 - (x^3 + ax + b) \quad (6.137)$$

When is 0 a critical value and when is it a regular value?

The inverse image of 0 is always nonempty and is known as an *elliptic curve*. Note that

$$dF = (-(3x^2 + a), 2y) \quad (6.138)$$

By the preimage theorem, if $dF \neq 0$ then the curve is nonsingular. Let us find the criterion for 0 to be a critical value of F . If $dF = 0$ then $y = 0$ and therefore any preimage of $F = 0$

must be $(0, \theta)$ where θ is a real root of the cubic polynomial $x^3 + ax + b$. But for $dF = 0$ we also must have $3\theta^2 + a = 0$. Thus, we must have a simultaneous solution of:

$$\begin{aligned}\theta^3 + a\theta + b &= 0 \\ 3\theta^2 + a &= 0\end{aligned}\tag{6.139}$$

We can solve for a and b in terms of this hypothetical θ to find:

$$\begin{aligned}a &= -3\theta^2 \\ b &= -2\theta^3\end{aligned}\tag{6.140}$$

which implies the necessary condition:

$$\Delta = 4a^3 + 27b^2 = 0\tag{6.141}$$

In fact, this is necessary and sufficient for 0 to be a critical value of F :

The quantity Δ is a famous quantity associated with a cubic polynomial $p(x)$. Such a polynomial always has three complex roots e_1, e_2, e_3 , and if we shift x so that the coefficient of x^2 vanishes then we can write:

$$p(x) = x^3 + ax + b = (x - e_1)(x - e_2)(x - e_3)\tag{6.142}$$

and hence

$$\begin{aligned}0 &= e_1 + e_2 + e_3 \\ a &= e_1e_2 + e_1e_3 + e_2e_3 = -(e_1^2 + e_2^2 + e_1e_2) \\ b &= -e_1e_2e_3 = e_1e_2(e_1 + e_2)\end{aligned}\tag{6.143}$$

A simple computation shows that

$$\Delta := 4a^3 + 27b^2 = -(e_1 - e_2)^2(e_1 - e_3)^2(e_2 - e_3)^2\tag{6.144}$$

and therefore the vanishing of Δ means that two roots coincide. But two roots coincide iff $p(x) = 0$ and $\frac{d}{dx}p(x) = 0$ have a common solution.

It is interesting to see how the nature of the level set $F^{-1}(0)$ changes as Δ moves through zero. From the above formulae we see that

1. $p(x)$ has three real roots iff $\Delta \leq 0$.
2. The three roots are distinct iff $\Delta < 0$.
3. If $\Delta = 0$ but $f < 0$ there are two distinct real roots (one of which is double).
4. The three roots coincide $e_1 = e_2 = e_3 = 0$ iff $\Delta = 0$ and $f = 0$.

See Figure 33 above.

Remarks

1. It is natural to ask whether every submanifold $N \subset M$ can be described globally as the solution set of some system of equations. In general that cannot be done. The obstruction is the nontriviality of the *normal bundle* of N in M . Indeed, N can be globally expressed as a solution set of some equations iff the normal bundle of N is trivializable.

2. Viewing submanifolds locally as defined by equations is very useful also in discussing infinite-dimensional manifolds. For example, in field theory, there might be an infinite-dimensional space of fields ϕ defined on a spatial domain, say $\phi \in \text{Map}(\mathcal{S}, X)$. The Hamiltonian, or energy functional H is a map $H : \text{Map}(\mathcal{S}, X) \rightarrow \mathbb{R}$ taking $\phi \rightarrow H[\phi] \in \mathbb{R}$. The level sets $H^{-1}(E)$ are the field configurations of constant energy and so long as $dH \neq 0$ that set of field configurations will be an infinite-dimensional submanifold of codimension one. Note in particular that while the RHS of (6.119) does not make sense in infinite-dimensions this local view of submanifolds does give a good definition of codimension for infinite-dimensional submanifolds.
3. Similarly, the implicit function theorem generalizes to infinite dimensions.
4. If we consider the space of all differentiable functions $f : M \rightarrow \mathbb{R}^\ell$ and $\dim M \geq \ell$ then the condition that df has rank ℓ is an *open* condition. The subspace of functions with a critical point is a complicated and interesting subspace in the space of all functions. It is generically of real codimension one, and hence splits the space $\text{Map}(M, \mathbb{R}^\ell)$ into open domains. The study of the critical locus is the subject of *singularity theory* or *catastrophe theory*.

Exercise

Consider $F : \mathbb{R}^3 \rightarrow \mathbb{R}$ given by

$$F(x, y, z) = z^2 - x^2 - y^2 \tag{6.145}$$

Show that 0 is not a regular value of F and draw $F^{-1}(0)$. Identify the critical points and regular points in the level set.

6.9 Lie groups

A very important concept in physics is that of a *Lie group*. This is a group which is also a manifold, such that the group and manifold structures interact nicely. Thus, the maps $G \times G \rightarrow G$ given by multiplication and $G \rightarrow G$ given by inversion are smooth maps of manifolds.

We can apply the Preimage Theorem of Section §6.8.3 to prove that the classical matrix groups are in fact Lie groups. First we prove they are all manifolds:

Example 1: $GL(n, \mathbb{R}) \subset \mathbb{R}^{n^2}$ and $GL(n, \mathbb{C}) \subset \mathbb{C}^{n^2} \cong \mathbb{R}^{2n^2}$ are both manifolds. The coordinates are the matrix elements. The inverse image $\det^{-1}(0)$ is a closed subset of $M_n(\kappa)$, with $\kappa = \mathbb{R}, \mathbb{C}$ and hence any invertible matrix has an open neighborhood of invertible matrices. The transition functions are the identity. The tangent space at any point $A \in GL(n, \kappa)$ is isomorphic to the vector space $M_n(\kappa)$ of $n \times n$ matrices over κ . The group operation of multiplication is polynomial in the matrix elements and hence

certainly a $\mathcal{C}^\infty$ function. The group operation of inversion is a rational function of the matrix elements and is also $\mathcal{C}^\infty$ on $GL(n, \kappa)$.

Example 2: Now, for $SL(n, \kappa)$ consider $f : M_n(\kappa) \rightarrow \kappa$ defined by $f(A) := \det A - 1$. We claim that $0 \in \kappa$ is a regular value of f . Indeed, if A is invertible then for any

$$M \in T_A M_n(\kappa) \cong M_n(\kappa) \quad (6.146)$$

we have

$$df_A(M) = \det A \operatorname{Tr}(A^{-1}M) \quad (6.147)$$

This is usually written as the (very useful) identity ³⁶

$$\delta \log \det A = \operatorname{Tr}(A^{-1} \delta A) \quad (6.148)$$

for A invertible. When A is invertible the kernel of df_A is the linear subspace of $n \times n$ matrices M such that $A^{-1}M$ is traceless, which is linearly equivalent to the linear subspace of traceless matrices, and therefore has dimension (over κ) equal to $n - 1$. Therefore the rank of df_A is 1, and $f = \det$ is a submersion. So the inverse image is a manifold.

Example 3: For $O(n; \kappa)$ we define $f : M_n(\kappa) \rightarrow S_n(\kappa)$ where $S_n(\kappa) \cong \kappa^{\frac{1}{2}n(n+1)}$ is the vector space over κ of $n \times n$ symmetric matrices. We take f to be

$$f(A) = AA^{tr} - 1 \quad (6.149)$$

Then $O(n) = f^{-1}(0)$. We aim to show it is a manifold. Note that df_A is a linear operator $M_n(\kappa) \rightarrow S_n(\kappa)$. It is just

$$df_A(M) = MA^{tr} + AM^{tr} \quad (6.150)$$

Therefore $\ker(df_A)$ is the linear subspace of $M_n(\kappa)$ of matrices such that MA^{tr} is anti-symmetric. When A is invertible this subspace is isomorphic to the linear subspace of anti-symmetric matrices and hence has dimension $\frac{1}{2}n(n-1)$. It follows that 0 is a regular value of f and $O(n, \kappa)$ is a manifold.

Example 4: For $Sp(2n; \kappa)$ we define $f : M_{2n}(\kappa) \rightarrow A_{2n}(\kappa)$ where $A_{2n}(\kappa)$ is the set of $(2n) \times (2n)$ matrices over κ such that (Jm) is antisymmetric. This is isomorphic to the vector space over κ of dimension $\frac{1}{2}(2n)(2n-1) = n(2n-1)$. Now we take f to be

$$f(A) = AJA^{tr}J^{tr} - 1 \quad (6.151)$$

so that $Sp(2n; \kappa) = f^{-1}(0)$. Again we claim that 0 is a regular value of f . Now df_A is the linear operator $M_{2n}(\kappa) \rightarrow A_{2n}(\kappa)$. It is just

$$df_A(M) = MJ A^{tr} J^{tr} + AJ M^{tr} J^{tr} \quad (6.152)$$

Therefore $\ker(df_A)$ is the linear subspace of $M_{2n}(\kappa)$ of matrices such that MJA^{tr} is symmetric. When A is invertible this subspace is isomorphic to the linear subspace of symmetric

³⁶For a proof see the Linear Algebra User's Manual, ch. 3.

matrices and hence has dimension $\frac{1}{2}2n(2n+1)$, which is complementary to the dimension of the image $\frac{1}{2}(2n)(2n-1)$ and hence df_A is surjective. It follows that 0 is a regular value of f and $Sp(2n; \kappa)$ is a manifold.

Example 5: Finally, for $U(n)$ consider $f : M_n(\mathbb{C}) \rightarrow \mathcal{H}_n$ where $\mathcal{H}_n$ is the *real* vector space of $n \times n$ Hermitian matrices in $M_n(\mathbb{C})$. This has real dimension $n + 2 \times \frac{1}{2}n(n-1) = n^2$. We now take $f(A) = AA^\dagger - 1$. Then

$$df_A(M) = MA^\dagger + AM^\dagger \quad (6.153)$$

When A is invertible the kernel is the subspace of $M_n(\mathbb{C})$ of matrices such that $MA^\dagger$ is anti-hermitian. This is again a real vector space of real dimension $n + 2 \times \frac{1}{2}n(n-1) = n^2$. Since $M_n(\mathbb{C})$ is a real vector space of real dimension $2n^2$ it follows that df_A is surjective and hence 0 is a regular value of f . Therefore $U(n)$ is a manifold.

Example 6: It is useful to combine the previous two examples and define the Lie group:

$$USp(2n) := U(2n) \cap Sp(2n, \mathbb{C}) \quad (6.154)$$

Now we have a map: $f : M_{2n}(\mathbb{C}) \rightarrow A_{2n}(\mathbb{C}) \oplus \mathcal{H}_{2n}$ defined by taking the direct sum. Again, one must check that the *real linear* map df_A at an invertible matrix in the preimage of 0 has a kernel of the correct dimension.³⁷

All the examples above are submanifolds of $GL(n, \kappa)$. It follows from the exercise of Section §6.8.2 that the group operations of multiplication and inversion are $\mathcal{C}^\infty$ maps. This concludes the argument that the above examples are all Lie groups.

6.9.1 Lie algebras of Lie groups

In general, the *Lie algebra* of a Lie group G is defined, as a vector space, to be the tangent space at the identity:

$$Lie(G) := T_1G \quad (6.155)$$

The reason for the term “algebra” will be explained below.

Our proof above that the classical matrix groups are manifolds also leads nicely to an immediate computation of the Lie algebras of these groups.

1. $GL(n, \kappa)$:

$$\mathfrak{gl}(n; \kappa) := T_1GL(n, \kappa) \cong M_n(\kappa) \quad (6.156)$$

³⁷We have not covered quaternions yet, but a superior viewpoint is to view $USp(2n)$ as the group of $n \times n$ unitary matrices over the quaternions. In this viewpoint we should define $f : M_n(\mathbb{H}) \rightarrow \mathcal{H}_n(\mathbb{H})$ where $\mathcal{H}_n(\mathbb{H})$ is the space of $n \times n$ quaternionic Hermitian matrices. The above arguments work in the same way: $\dim_{\mathbb{R}} M_n(\mathbb{H}) = 4n^2$, while $\dim_{\mathbb{R}} \mathcal{H}_n(\mathbb{H}) = n + 4 \times \frac{1}{2}n(n-1) = 2n^2 + n$. As before, the kernel of df_A , for A invertible is the space of $n \times n$ quaternionic-antihermitian matrices. This has real dimension $3n + 4 \times \frac{1}{2}n(n-1) = 2n^2 + n$. (The $3n$ is there because one can have an arbitrary imaginary quaternion on the diagonal.) Therefore 0 is a regular value, and $USp(2n)$ is a manifold. Many authors denote this group simply as $Sp(n)$.

2. $SL(n, \kappa)$:

$$\mathfrak{sl}(n; \kappa) := T_1 SL(n, \kappa) \cong \{M \in M_n(\kappa) | \text{Tr}(M) = 0\} \quad (6.157)$$

3. $O(n, \kappa)$:

$$\mathfrak{o}(n; \kappa) = \mathfrak{so}(n; \kappa) := T_1 O(n, \kappa) \cong \{M \in M_n(\kappa) | M^{tr} = -M\} \quad (6.158)$$

4. $Sp(2n, \kappa)$:

$$\mathfrak{sp}(2n; \kappa) := T_1 Sp(2n, \kappa) \cong \{M \in M_n(\kappa) | (MJ)^{tr} = +MJ\} \quad (6.159)$$

5. $U(n)$:

$$\mathfrak{u}(n) := T_1 U(n) \cong \{M \in M_n(\mathbb{C}) | M^\dagger = -M\} \quad (6.160)$$

6. $SU(n)$:

$$\mathfrak{su}(n) := T_1 SU(n) \cong \{M \in M_n(\mathbb{C}) | M^\dagger = -M \quad \& \quad \text{Tr}(M) = 0\} \quad (6.161)$$

7. $USp(2n)$:

$$\begin{aligned} \mathfrak{usp}(2n) &:= T_1 USp(2n) \cong \{M \in M_{2n}(\mathbb{C}) | M^\dagger = -M \quad \& \quad (MJ)^{tr} = MJ\} \\ &\cong \{M \in M_n(\mathbb{H}) | M^\dagger = -M\} \end{aligned} \quad (6.162)$$

In each case, given a matrix $V \in T_1 G \subset M_N(\kappa)$ we can form the family of group elements

$$g_V(t) = \exp[tV] := \sum_{n=0}^{\infty} \frac{(tV)^n}{n!} \quad (6.163)$$

Note that for $t \in \mathbb{R}$ these elements form a subgroup of G :

$$g_V(t_1)g_V(t_2) = g_V(t_1 + t_2) \quad (6.164)$$

Indeed $t \mapsto g_V(t)$ is a homomorphism of $\mathbb{R} \rightarrow G$. It is a good exercise to check, from the defining relations of $T_1 G$ above that the exponentiated matrix indeed satisfies the defining relations of the group. Thus, for example, one should check that if M is anti-Hermitian, i.e. if $M^\dagger = -M$ then $\exp(M)$ is unitary.

An important property of the tangent spaces $T_1 G$ for the various groups above is that:

If $V_1, V_2 \in T_1 G$ then the matrix commutator $[V_1, V_2]$ is also in $T_1 G$.

This can be verified by directly checking each case. For example, in the case of $\mathfrak{so}(n, \kappa)$, if V_1, V_2 are antisymmetric matrices over κ then neither $V_1 V_2$, nor $V_2 V_1$ is antisymmetric, but $[V_1, V_2]$ is. The reader should check the other cases in this way. Nevertheless, this fact also follows from more general principles, and that is important because as we will see not every Lie group is a classical matrix group. In fact, it is not true that every finite-dimensional Lie group is a subgroup of $GL(N, \mathbb{R})$ for some N .³⁸ Given V_1, V_2 we can consider the path through $g = 1$ at $t = 0$ given by the group commutator:

$$\lambda(t) = [g_{V_1}(\sqrt{t}), g_{V_2}(\sqrt{t})] \quad (6.165)$$

³⁸A counterexample is the metaplectic group, a group which arises as a central extension of the symplectic group when one tries to implement symplectic transformations on the quantum mechanics of a system of free particles.

Now, one can show that for t_1, t_2 small we have ³⁹

$$g_{V_1}(t_1)g_{V_2}(t_2) = \exp[t_1V_1 + t_2V_2 + \frac{1}{2}t_1t_2[V_1, V_2] + \mathcal{O}(t_1^a t_2^b)] \quad (6.166)$$

where the higher order terms have $a + b > 2$, and therefore the tangent vector to the path through $\lambda(t)$ is the matrix commutator.

Just based on group theory and manifold theory, therefore, one can deduce the following about the vector space over κ defined by $\mathfrak{g} = T_1G$:

1. There is an antisymmetric, bilinear multiplication

$$[\cdot, \cdot] : \mathfrak{g} \times \mathfrak{g} \rightarrow \mathfrak{g} \quad (6.167)$$

2. For all $v_1, v_2, v_3 \in \mathfrak{g}$ we have the Jacobi identity:

$$[v_1, [v_2, v_3]] + cycl. = 0 \quad (6.168)$$

simply because this is a property of matrix commutators.

These are the defining properties of a Lie algebra. Here is the formal definition:

Definition A *Lie algebra* is a vector space $\mathfrak{g}$ over a field κ such that there is a *Lie bracket*, $\mathfrak{g} \times \mathfrak{g} \rightarrow \mathfrak{g}$ such that for all vectors $v_1, v_2, v_3 \in \mathfrak{g}$ and scalars $\alpha, \beta \in \kappa$:

1. $[v_1, v_2] = -[v_2, v_1]$.
2. $[\alpha v_1 + \beta v_2, v_3] = \alpha[v_1, v_3] + \beta[v_2, v_3]$
3. The *Jacobi identity* holds:

$$[v_1, [v_2, v_3]] + [v_2, [v_3, v_1]] + [v_3, [v_1, v_2]] = 0 \quad (6.169)$$

Remarks

1. As indicated above, one can show that for any Lie group the tangent space $\mathfrak{g} = T_1G$ is a Lie algebra. The Jacobi identity follows from the associativity of group multiplication.
2. As we observed above, the vector space $\text{Vect}(M)$ over $\mathbb{R}$ of all $\mathcal{C}^\infty$ vector fields on a manifold is a Lie algebra (of infinite dimension).
3. Every Lie group has a corresponding Lie algebra, but the converse is not always the case. A notable example is the complexified Lie algebra of vector fields on the circle.

³⁹The full series is known, and is occasionally useful. It is known as the Baker-Campbell-Hausdorff formula. See Chapter *** . For present purposes see the exercise below.

4. It is also worth noting that every $V \in T_1G$ is associated with two vector fields defined on all of G . If $g_0 \in G$ then $L(g_0) : G \rightarrow G$ is the action of left-multiplication defined by $g \mapsto g_0g$ and similarly $R(g_0) : g \mapsto gg_0$. Then $L(g_0)_* : T_1G \rightarrow T_{g_0}G$, and so if $V \in T_1G$ then $(\xi_L(V))_{g_0} := L(g_0)_*(V)$ defines a vector field on all of G called a *left-invariant vector field* because $L(g_1)_*(\xi(V)_{g_2}) = \xi(V)_{g_1g_2}$. For $G = GL(n, \mathbb{R})$ we can use the matrix elements g_{ij} as coordinates on the group and one can show that for $V = e_{ij}$ (the matrix unit in $M_n(\mathbb{R})$) we have

$$\xi(V)_g = \left(g^{tr} \frac{\partial}{\partial g} \right)_{ij} = g_{ki} \frac{\partial}{\partial g_{kj}} \quad (6.170)$$

Similarly, one can define right-invariant vector fields using $R(g)_*$. The tangent to the path $g_V(t)$ at $t = t_0$ is the left invariant vector field $\xi_L(V)$ at $g_V(t_0)$. (See the Section on group actions below.)

Exercise *BCH to lowest nontrivial order*

Prove (6.166):

- a.) Show that we need only consider terms

$$g_{V_1}(t_1)g_{V_2}(t_2) = 1 + t_1V_1 + t_2V_2 + t_1t_2V_1V_2 + \cdots \quad (6.171)$$

- b.) Now take the logarithm of the above expression and expand to the required order.
-

Exercise

We defined an algebra $\mathcal{A}$ as a vector space over a field κ with a bilinear distributive product $\mathcal{A} \times \mathcal{A} \rightarrow \mathcal{A}$.

An algebra is said to be *associative* if $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ for all vectors $a, b, c \in \mathcal{A}$.

Is a Lie algebra an associative algebra?

Exercise

Let T^a be a vector space basis for a Lie algebra $\mathfrak{g}$. Then there must be elements $f_c^{ab} \in \kappa$ - known as *structure constants* - such that

$$[T^a, T^b] = f_c^{ab} T^c \quad (6.172)$$

- a.) Show that $f_c^{ab} = -f_c^{ba}$

- b.) Write out the identity on the f_c^{ab} implied by the Jacobi identity.
-

6.9.2 Remarks on the classification of Lie groups

As with the case of groups, we cannot hope to classify all Lie groups. We need to put some simplifying criteria.

The manifolds $SU(n)$, $SO(n)$ and $USp(2n) := U(2n) \cap Sp(2n, \mathbb{C})$ are all compact. Can we classify all compact groups? No: Any finite group is a Lie group. (Manifolds can have dimension zero!) So, what if we insist the dimension is positive? Still no: Given any finite group F and any Lie group G then $G \times F$ is another Lie group. What about compact *connected* groups? This is still hard.

A good criterion turns out to be to classify *connected*, *simply connected*, *simple Lie groups*. A Lie group is said to be a *simple Lie group* if it has no nontrivial connected normal subgroups.⁴⁰

Thus, for example, $U(n)$ is connected and compact, but is not a simple Lie group because the central subgroup

$$\mathcal{Z}(U(n)) = \{u = z1_{n \times n} | z \in \mathbb{C} \quad \& \quad |z| = 1\} \cong U(1) \quad (6.173)$$

is a nontrivial connected normal subgroup. However, one can show that $SU(n)$ is compact, connected, simple, and simply connected. Similarly, one can show that $USp(2n)$ has all these properties. Are there others?

$SO(n)$ is compact, connected, and simple, but not quite simply connected. In fact, for $n > 2$, $\pi_1(SO(n)) \cong \mathbb{Z}_2$. We don't have the tools to prove this, but there is a standard "coffee cup proof" for the case of $n = 3$ which will do for the moment. See Figure 34.

The universal cover (see Section §13.7 below) of $SO(n)$ is still a Lie group and is known as the *spin group* and will be denoted as $\text{Spin}(n)$. Since $\pi_1(SO(n)) \cong \mathbb{Z}_2$ it is a 2-fold cover and is therefore connected simply connected and compact and fits in the exact sequence:

$$1 \rightarrow \mathbb{Z}_2 \rightarrow \text{Spin}(n) \rightarrow SO(n) \rightarrow 1 \quad (6.174)$$

The group $\text{Spin}(n)$ is best constructed directly using Clifford algebras. See Chapter ****.

Are there more compact connected simply connected Lie groups? The classification of Lie groups was begun by Wilhelm Killing and Friedrich Engel in the late 19th century and for the case of compact connected simply connected Lie groups it was definitively completed by Elie Cartan in his PhD thesis of 1894. The essential technique is to reduce the problem to an algebraic problem using the Lie algebra of the Lie group. As in the classification of finite simple groups, there is a list of infinite series of "easy" examples and a finite list of exceptional examples. The full list of compact, simple, connected, and simply connected Lie groups is:

⁴⁰Unfortunately, there is some variation in the literature about the definition of a simple Lie group. There is no disagreement about the definition of a simple Lie algebra: A *simple Lie algebra* is one with no nontrivial Lie algebra ideals. One definition of a simple Lie group is that it has a simple Lie algebra. We would then have to admit $\mathbb{R}$, $U(1)$, $O(n)$, ... as simple Lie groups.



(a)



(b)

Figure 34: The coffee cup proof that $\pi_1(SO(3)) \cong \mathbb{Z}/2\mathbb{Z}$. In part (a) imagine a line of frames for $\mathbb{R}^3$ stretching along the arm from shoulder to cup. Now, twisting the arm as in (b) the frames along the arm are rotated, but the final orientation of the cup is the same as in (a). Therefore, the frames describe a closed loop in $SO(3)$. You know in your bones that the loop is nontrivial! Moreover, composing the motion twice gives back the identity. A respectable mathematical proof uses the construction of $\text{Spin}(n)$ in terms of Clifford algebras.

Name	Real Dimension	Lie algebra	Cartan symbol	Range of n
$SU(n+1)$	$n(n+2)$	$\mathfrak{su}(n+1)$	A_n	$n \geq 1$
$\text{Spin}(2n+1)$	$n(2n+1)$	$\mathfrak{so}(2n+1)$	B_n	$n \geq 1$
$USp(2n)$	$n(2n+1)$	$\mathfrak{usp}(2n)$	C_n	$n \geq 1$
$\text{Spin}(2n)$	$n(2n-1)$	$\mathfrak{so}(2n)$	D_n	$n \geq 3$
G_2	14	$\mathfrak{g}_2$	G_2	
F_4	52	$\mathfrak{f}_4$	F_4	
E_6	78	$\mathfrak{e}_6$	E_6	
E_7	133	$\mathfrak{e}_7$	E_7	
E_8	248	$\mathfrak{e}_8$	E_8	

A few remarks are in order:

1. First, the structure of $\text{Spin}(n)$ is very different in the Cartan-Killing classification for n even and odd, and hence they are viewed as two different cases.
2. The group in the above list are all non-isomorphic with the (very significant!) exception of a few examples in low dimension. These are:
 - $\text{Spin}(2) \cong U(1)$. (This group is abelian, compact, and not simply connected. According to some definitions it would be simple.)
 - $\text{Spin}(3) \cong SU(2) \cong USp(2)$
 - $\text{Spin}(4) \cong SU(2) \times SU(2)$ (Note this is a not simple group!)
 - $\text{Spin}(5) \cong USp(4)$
 - $\text{Spin}(6) \cong SU(4)$

These isomorphisms are proved using Clifford algebras (note they all involve spin groups).

3. If a compact Lie group has a compact universal cover which is a product of simple Lie groups it is *semi-simple*. Thus, $SU(2) \times SU(2)$ and $SO(4)$ are semisimple.

Exercise *Lie groups from indefinite forms*

In general, if Q is a quadratic form on a vector space over κ then $O(Q)$ is the automorphism group of the quadratic form. If h is an Hermitian form on a complex vector space then $U(h)$ is the group of complex linear automorphisms of h .

Consider the particular example of the matrix:

$$\eta_{p,q} = \begin{pmatrix} +\mathbf{1}_{p \times p} & 0 \\ 0 & -\mathbf{1}_{q \times q} \end{pmatrix} \quad (6.175)$$

where p, q are positive integers. Let $O(p, q)$ be the subgroup of $A \in GL(n; \mathbb{R})$ with $n = p + q$ such that $A\eta_{p,q}A^{tr} = \eta_{p,q}$.

a.) Show that $O(p, q)$ is a Lie group.

Now let $U(p, q)$ be the subgroup of $A \in GL(n; \mathbb{C})$ such that $A\eta_{p,q}A^\dagger = \eta_{p,q}$.

b.) Show that $U(p, q)$ is a Lie group.

6.10 Transversality

Suppose that Σ_1 and Σ_2 are two submanifolds of a manifold M . A natural question we can ask is whether or not they will intersect. Here the concept of codimension becomes quite useful.

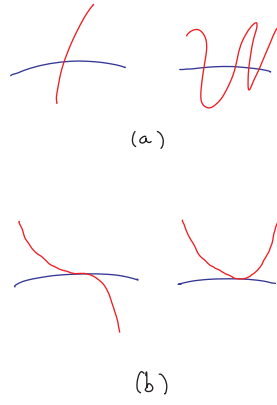


Figure 35: (a): Transversal intersections in two dimensions. (b): Nontransversal intersections in two dimension. Note that the codimension rule fails in some of the examples (b).

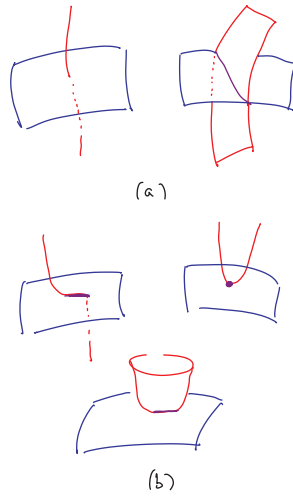


Figure 36: (a): Transversal intersections in three dimensions. (b): Nontransversal intersections in three dimension. Note that the codimension rule fails in some of the examples (b).

Suppose $\text{cod}\Sigma_1 = \ell_1$ and $\text{cod}\Sigma_2 = \ell_2$. Then, locally, Σ_i are defined by the level sets of ℓ_i different functions, say $(f_1, \dots, f_{\ell_1}) = c$ and $(g_1, \dots, g_{\ell_2}) = c'$. If these functions are independent then in a neighborhood where they intersect we expect the intersection to be a new submanifold of codimension $\ell_1 + \ell_2$. That is, we expect *codimension should add under intersection*.

To make this precise we need a notion of transversal intersection:

Definition: Two submanifolds $\Sigma_1, \Sigma_2 \subset M$ are said to *intersect transversally* at a point

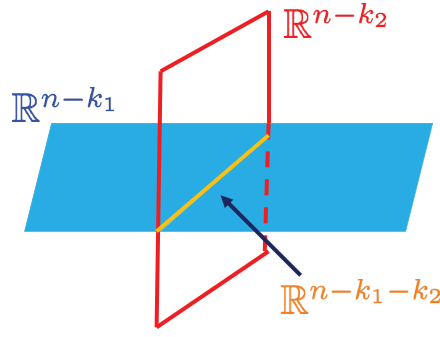


Figure 37: Illustrating the local picture of a transversal intersection of two submanifolds of codimension k_1 and k_2 in a manifold of dimension n .

$p \in \Sigma_1 \cap \Sigma_2$ if

$$T_p \Sigma_1 + T_p \Sigma_2 = T_p M \quad (6.176)$$

where on the LHS we mean the vector space generated by the sum of the vectors in $T_p \Sigma_1$ and $T_p \Sigma_2$. In particular, it is not a direct sum, in general.

Examples: Examples of transversal and nontransversal intersections in two and three dimensions are shown in Figure 35 and Figure 36.

Now we have the important

Theorem: If two submanifolds $\Sigma_1, \Sigma_2 \subset M$ intersect transversally (i.e. every $p \in \Sigma_1 \cap \Sigma_2$ is a transversal intersection) then $\Sigma_1 \cap \Sigma_2$ is a submanifold of codimension

$$\text{cod}(\Sigma_1 \cap \Sigma_2) = \text{cod}(\Sigma_1) + \text{cod}(\Sigma_2) \quad (6.177)$$

Equivalently (in finite dimensions):

$$\dim(\Sigma_1 \cap \Sigma_2) = \dim(\Sigma_1) + \dim(\Sigma_2) - \dim M \quad (6.178)$$

Moreover,

$$T_p(\Sigma_1 \cap \Sigma_2) = T_p \Sigma_1 \cap T_p \Sigma_2. \quad (6.179)$$

For a proof e.g. Guilleman and Pollack, *Differential Topology*. The essential idea is that the submanifolds are locally defined by collections of independent functions. See Figure 37. Sometimes a transversal intersection is denoted $\Sigma_1 \pitchfork \Sigma_2$.

Equations (6.177) and (6.178) can be taken as a useful rule of thumb: Note that the codimension of any submanifold of M must be $\leq n = \dim M$. (A codimension zero submanifold is a discrete set of points.) So, we expect, “in general,” that two submanifolds Σ_1 and Σ_2 with $\text{cod}(\Sigma_1) + \text{cod}(\Sigma_2) \leq n$ will have a nonzero intersection. Equivalently, if

$$\dim(\Sigma_1) + \dim(\Sigma_2) \geq \dim M \quad (6.180)$$

we expect, “in general,” that they will intersect.

As simple examples, note that two straight lines in the plane in general intersect, and a straight line and a plane in $\mathbb{R}^3$ in general intersect.

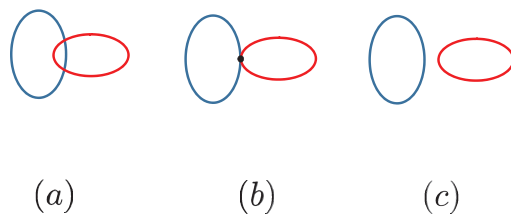


Figure 38: The blue and red submanifolds of the plane might or might intersect. In (a) they intersect transversally. In (b) they intersect nontransversally.

Of course, it is extremely easy to produce lots of counterexamples to the rule of thumb:

1. Two points in M each have codimension n so the sum of codimensions is $2n > n$. Indeed, two randomly chosen points in M will be different and hence will not intersect. But if they are the same point then the submanifolds intersect. Note that (6.177) is maximally violated!
2. One dimensional submanifolds manifolds in $\mathbb{R}^2$ have a sum of codimensions which is 2, so in general we expect them to intersect in a codimension two submanifold. That is, we expect them to intersect in points. But of course it is easy to draw examples where they do not intersect as in Figure 38.
3. One dimensional submanifolds manifolds in $\mathbb{R}^3$ have a sum of codimensions which is $4 > 3$, and indeed do not in general intersect. But of course, we can also have intersecting one-dimensional submanifolds in $\mathbb{R}^3$.

The notion of “generic” can be made more precise:

A property P of a function f is said to be stable if for any homotopy f_t with $t \in (a, b)$, $a < 0 < b$, and $f_0 = f$, there is a sufficiently small ϵ such that the property P is true of f_t for all $t \in (-\epsilon, \epsilon)$.

Given a transverse intersection of submanifolds the property of intersecting is a stable notion. (The function in question is the immersion of the respective submanifolds.)

Some important properties which are stable are

1. f is a diffeomorphism.
2. f is an immersion, submersion, or embedding

Example: An example which is popular in string theory is the following: Let us consider the intersection of two affine planes in $\mathbb{R}^N$. Suppose they have dimensions p and $N - p$. Without loss of generality we can choose coordinates $(x^1, \dots, x^N)$ on $\mathbb{R}^N$ so that the p -dimensional plane Σ_1 is given by

$$x^{p+1} = x^{p+2} = \dots = x^N = 0 \quad (6.181)$$

Now, the general $(N - p)$ -dimensional affine plane Σ_2 will be determined by p equations of the form

$$\sum_{j=1}^N A_{aj} x^j = L_a \quad 1 \leq a \leq p \quad (6.182)$$

where A is an $(N - p) \times N$ matrix. We compute the intersection $\Sigma_1 \cap \Sigma_2$ by setting the final $(N - p)$ coordinates to zero in (6.183). Thus, the intersection consists of points $(x^1, \dots, x^p, 0, \dots, 0)$ such that

$$\sum_{b=1}^p A_{ab} x^b = L_a \quad 1 \leq a \leq p \quad (6.183)$$

Now, for a *generic* hyperplane, A_{ab} will be invertible and there will be a solution. Thus, the hyperplanes intersect in a unique point. However, if A_{ab} has rank less than p , and the vector L_a is in the image of A the intersection will be a hyperplane of positive dimension. If A_{aj} has rank less than p a generic perturbation will make it rank p . Thus, the higher-dimensional intersections are unstable, but the 0-dimensional intersections are stable.

Exercise

a.) Consider two 5-dimensional hyperplanes in 10-dimensional Euclidean space. Find configurations in which they intersect in hyperplanes of dimensions 0, 1, 2, 3, 4, 5.

b.) Consider hyperplanes of dimension p_1 and p_2 in $\mathbb{R}^N$. Write out conditions for them to intersect in hyperplanes of dimensions $\leq \text{Min}[p_1, p_2]$.

6.10.1 Relative Transversality

TO BE WRITTEN.

6.11 Intersection Numbers

A case of particular importance is the case when Σ_1 and Σ_2 have complementary dimensions in M , that is,

$$\dim \Sigma_1 + \dim \Sigma_2 = \dim M \quad (6.184)$$

Or, equivalently

$$\text{cod} \Sigma_1 + \text{cod} \Sigma_2 = \dim M \quad (6.185)$$

In this case the rule of thumb suggests they will generically intersect in a collection of discrete points. There is a way to associate a deformation invariant quantity associated with such configurations known as the *intersection number*.

As Figure 38 shows, just counting the number of points of intersection is not going to be deformation invariant. However, one can prove ⁴¹ that *the number of points modulo two* is invariant, provided the intersections are transverse. See Figure 38. In (a) and (c) the number of intersections modulo two is zero. In (b) it is one, but this is a nontransversal intersection. This invariant is called the mod-two intersection number.

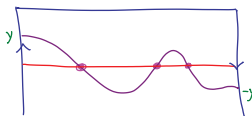


Figure 39: The purple curve is a perturbation of the central red curve. The self-intersection modulo two is $1 \bmod 2$.

Example: Consider the central circle of a cylinder. The self-intersection is zero. Now consider the central circle of a Mobius strip. The self-intersection (modulo two) is 1. This again reflects the nontrivial topology of the strip. See Figure 39.

If, moreover, M is orientable and Σ_1 and Σ_2 are also orientable then we can go further and assign an integer, not just an integer modulo two.

In order to do this note that if Σ_1 and Σ_2 intersect transversally at $p \in M$ then we have a direct sum

$$T_p \Sigma_1 \oplus T_p \Sigma_2 \cong T_p M \quad (6.186)$$

Consequently, if Σ_1 , Σ_2 and M are all oriented then any oriented basis $\{v_1, \dots, v_{k_1}\}$ for $T_p \Sigma_1$ and $\{w_1, \dots, w_{k_2}\}$ for $T_p \Sigma_2$ can be used to produce a basis

$$\{v_1, \dots, v_{k_1}, w_1, \dots, w_{k_2}\} \quad (6.187)$$

for $T_p M$. We say the local intersection number at p , denoted $\iota_p(\Sigma_1, \Sigma_2)$ is $+1$ if the orientation agrees with M and is -1 if it disagrees.

It is not difficult to see that $\iota_p(\Sigma_1, \Sigma_2) = \pm \iota_p(\Sigma_2, \Sigma_1)$ and the pairing is antisymmetric iff $\text{cod} \Sigma_1$ and $\text{cod} \Sigma_2$ are odd.

⁴¹See, for example, Guilleman and Pollack, *Differential Topology*

If $\Sigma_1 \pitchfork \Sigma_2$ intersect transversally and there are a finite number of intersection points then we set:

$$\iota(\Sigma_1, \Sigma_2) := \sum_{p \in \Sigma_1 \cap \Sigma_2} \iota_p(\Sigma_1, \Sigma_2) \quad (6.188)$$

This number is called the *oriented intersection number* and is a deformation invariant.

Remarks:

1. If $\Sigma \subset M$ is a submanifold of half the dimension of M then the *self-intersection* is obtained by making a generic perturbation of $\tilde{\Sigma}$ of Σ and computing the number of intersections of $\Sigma \pitchfork \tilde{\Sigma}$ modulo 2. If Σ and M are both oriented then we compute $\iota(\Sigma, \tilde{\Sigma})$.
2. One can show that the Euler character of a manifold M can be the self-intersection of the diagonal $\Delta = \{(p, p) | p \in M\} \subset M \times M$.

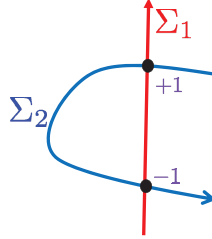


Figure 40: Two oriented one-dimensional manifolds intersect in the plane with standard orientation: A basis $\{e_1, e_2\}$ is in the standard orientation if $e_1 \times e_2$ points upward. The two intersection points make equal and opposite contributions to the intersection number $\iota(\Sigma_1, \Sigma_2)$. The intersections can clearly be removed by a smooth deformation of, say, Σ_2 .

6.11.1 The Whitney disk trick

Suppose we have two transversally intersecting submanifolds Σ_1 and Σ_2 in a manifold M . We can ask whether we can deform them continuously to make the smallest number of intersections. Ideally, the intersection number $\iota(\Sigma_1, \Sigma_2)$ should have no cancelling terms. For example, for oriented curves in the plane this can clearly be done. See Figure 40. How does this generalize to higher dimensions? Suppose we have the situation shown in Figure 41. The Whitney disk trick involves choosing two curves α, β connecting the intersections with opposite intersection number so that $\alpha \subset \Sigma_1$ and $\beta \subset \Sigma_2$, and so that they do not pass through any other intersections of Σ_1 and Σ_2 . Then, *provided*

$$\text{cod}(\Sigma_1) + \text{cod}(\Sigma_2) = \dim M \geq 5 \quad (6.189)$$

one can bound $\alpha \circ \beta$ by an *embedded* disk.

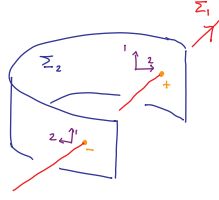


Figure 41: Generalizing the deformation eliminating canceling points to higher dimensions.

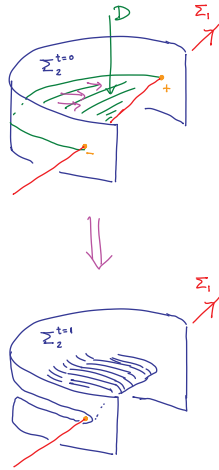


Figure 42: Generalizing the deformation eliminating canceling points to higher dimensions.

This should be plausible based on what we have proven: Recall that it was easy to prove the version of the Whitney embedding theorem where an n -manifold could be embedded into $\mathbb{R}^{2n+1}$. Here we have an extra constraint provided by the edge of the disk, so we can only expect to embed the disk in 5 or higher dimensions.

Once we have a nonself-intersecting disk we can use it to deform Σ_2 along the disk to make a homotopy Σ_2^t which removes the two intersections that were giving canceling contributions to $\iota(\Sigma_1, \Sigma_2)$. See Figure 42. Note that this suggests there might be an important difference between topology above and below four dimensions. We will come back to this point.

For more about this, and further references, see the very nice discussion by R. Kirby

6.11.2 Intersection pairing and homology theory

The intersection number plays an important role in *homology theory*. See Section **** for a formal definition of (co)homology theory. For our present purposes we for an abelian group A we define $H_p(M; A)$ to be a quotient of abelian groups Z_p/B_p . The abelian group Z_p generated by all closed p -dimensional submanifolds⁴² and the subgroup B_p is generated by submanifolds which are boundaries of $(p + 1)$ -dimensional submanifolds. The groups $H_p(M; \mathbb{Z}_2)$ are defined using all submanifolds while the groups $H_p(M; \mathbb{Z})$ are defined using only oriented submanifolds.

The deformation invariance of the intersection number is sufficiently strong that it descends to pairings on the homology groups:

$$H_p(M; \mathbb{Z}_2) \otimes H_{n-p}(M; \mathbb{Z}_2) \rightarrow \mathbb{Z}_2 \quad (6.190)$$

$$H_p(M; \mathbb{Z}) \otimes H_{n-p}(M; \mathbb{Z}) \rightarrow \mathbb{Z} \quad (6.191)$$

where in the latter case M should be compact and oriented.

Example 1: As an example consider $H_1(\Sigma; \mathbb{Z})$ for an oriented compact surface Σ of genus g . In this case $H_1(\Sigma; \mathbb{Z}) \cong \mathbb{Z}^{2g}$ and a standard basis can be represented by the one-dimensional submanifolds A_I, B^I , $I = 1, \dots, g$ shown in Figure 43. The intersection form with respect to this basis is just the standard matrix J used to define a symplectic transformation.

Example 2: A very significant example of the above pairing is the intersection pairing of $H_2(M; \mathbb{Z})$ when M is a compact oriented four-dimensional manifold. $H_2(M; \mathbb{Z})$ is an abelian group, and it can be shown to be finitely generated. Therefore there is an exact sequence

$$0 \rightarrow \text{Tors}(H_2(M; \mathbb{Z})) \rightarrow H_2(M; \mathbb{Z}) \rightarrow \mathbb{Z}^{b_2} \rightarrow 0 \quad (6.192)$$

The rank b_2 is called the *second Betti number*. Recall that $\text{Tors}(H_2(M; \mathbb{Z}))$ is the torsion subgroup. This is a finite abelian group. Recall that means that for any $[\Sigma] \in \text{Tors}(H_2(M; \mathbb{Z}))$ there is an integer n so that $n[\Sigma] = 0$. Since the intersection pairing is linear, it vanishes on the torsion subgroup and therefore passes to a quadratic form

$$Q : \mathbb{Z}^{b_2} \times \mathbb{Z}^{b_2} \rightarrow \mathbb{Z} \quad (6.193)$$

Choosing a basis, Q is represented by a symmetric integral matrix q . If we change basis then the matrix changes by $q \rightarrow gqg^{tr}$ where $g \in GL(b_2; \mathbb{Z})$. Moreover, a deep theorem called *Poincaré duality* ensures that the matrix q has determinant $\det(q) = \pm 1$.

⁴²This is inaccurate, but will serve for the present heuristic discussion. Actually, one wants to replace p -dimensional submanifolds by “ p -cycles.” These are formal linear combinations of continuous maps from the p -simplex into the space M such that a (suitably defined) boundary vanishes. They are considered up to additions of boundaries of $(p + 1)$ -dimensional simplices. It is a nontrivial problem to decide if a homology class can be represented by a submanifold and counterexamples exist in higher dimensions. This problem was addressed by R. Thom in “Quelques propri  s globales des var  t  s diff  rentiables.” The first example is a 7-dimensional homology class in a 10-manifold. See C. Bohr, B. Hanke, and D. Kotschick, <http://arxiv.org/pdf/math/0011178.pdf>.

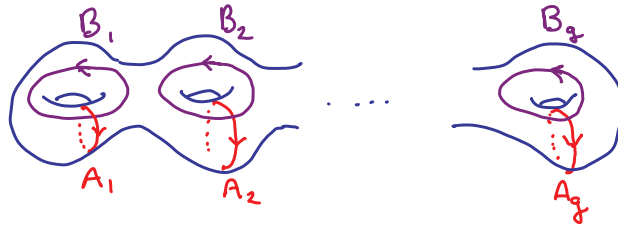


Figure 43: Standard A- and B- cycles on a genus g surface.

As a good example, consider $M = S^2 \times S^2$. Then $[\Sigma_1] = [S^2 \times pt]$ and $[\Sigma_2] = [pt \times S^2]$ generated $H_2(M; \mathbb{Z})$ and in this basis

$$Q = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad (6.194)$$

6.12 Linking

Closely related to the above case is the case when two closed submanifolds satisfy

$$\text{cod}(\Sigma_1) + \text{cod}(\Sigma_2) = \dim M + 1. \quad (6.195)$$

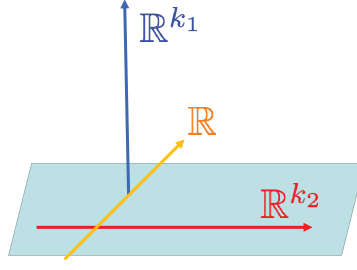


Figure 44: The blue and red submanifolds have one transverse direction in common and globally could link.

that is

$$\dim(\Sigma_1) + \dim(\Sigma_2) = \dim M - 1. \quad (6.196)$$

In this case we can sometimes define a *linking number*. The basic idea is that the local picture of two such submanifolds would be that shown in Figure 44.

For example, if $f_1 : S^{k_1} \rightarrow \mathbb{R}^{k_1+k_2+1}$ and $f_2 : S^{k_2} \rightarrow \mathbb{R}^{k_1+k_2+1}$ is a map of spheres then the images Σ_1, Σ_2 will generically link. This generalizes the linking of two knots in $\mathbb{R}^3$. We can define the linking number by extending f_1 to a map of a disk or f_2 to a map of a disk and computing the resulting intersection number:

$$L(\Sigma_1, \Sigma_2) := \iota(D_1, \Sigma_2) = \iota(\Sigma_1, D_2) \quad (6.197)$$

Note that this generalizes the Gauss linking number of Section *** above.

Example 1: One of the simplest topological field theories is “BF-theory,” which, in its simplest incarnation is a theory of abelian gauge potentials which are k -forms. So A is a k_1 -form on spacetime M with fieldstrength $F = dA$. Similarly, B is a k_2 -form with fieldstrength $G = dB$. Gauge transformations include $A \rightarrow A + d\lambda$ and $B \rightarrow B + d\mu$. Then, if M is oriented and of dimension $k_1 + k_2 + 1$ we can form the action

$$S \sim \int_M A dB \quad (6.198)$$

which is gauge invariant when M is closed. Gauge invariant observables can be associated to k_1 - and k_2 -dimensional submanifolds, respectively. They generalize Wilson lines:

$$W_{q_e}(\Sigma_1) := \exp i q_e \int_{\Sigma_1} A \quad W_{q_m}(\Sigma_2) := \exp i q_m \int_{\Sigma_2} B \quad (6.199)$$

In general these expressions only really make sense when Σ_1 and Σ_2 are boundaries of some other submanifold in M . In this very simple quantum field theory one can show that the correlation function is of the form

$$\exp[\kappa q_e q_m L(\Sigma_1, \Sigma_2)] \quad (6.200)$$

where κ is a constant that depends on how we normalize the action.

Example 2: The above ideas are often applied in discussions of “p-branes” in supergravity and string theories. In physics a “p-brane” is a p-space-dimensional extended object moving in spacetime. The worldvolume is $(p+1)$ -dimensional and is generally taken to be an embedded submanifold. In these theories there are generalizations of Maxwell theories with fieldstrengths which are totally antisymmetric tensors $F_{\mu_1 \dots \mu_n}$ of various ranks n . Thus they can be viewed as n -forms

$$F = \frac{1}{n!} F_{\mu_1 \dots \mu_n} dx^{\mu_1} \wedge \dots \wedge dx^{\mu_n} \quad (6.201)$$

In general, if a p -brane fills a submanifold of codimension r , so that $(p+1) + r = D$ is the dimension of spacetime, then there is a corresponding fieldstrength of degree $(r-1)$ so that the p -brane is a magnetic source:

$$dF = Q_p \prod_{i=1}^r \delta(f_i) df_i \quad (6.202)$$

where the brane worldvolume is locally defined by the r equations $f_1 = \dots = f_r = 0$ and Q_p is some way of parametrizing the charge of the brane with respect to this gauge field.

The generalization of Gauss’s law for measuring charge states that the charge of the p -brane can be measured by integrating F over a linking $(r-1) = D - p - 2$ sphere. Here D is the dimension of spacetime.

Now, in generalized Maxwell theories there is a generalized notion of electromagnetic duality. Two branes with p_1, p_2 are electro-magnetically dual if they are magnetic sources for Hodge-dual fieldstrengths: $F_1 = *F_2$. Here the only thing you need to know about Hodge $*$ is that it is a linear operator taking n -forms to $D - n$ forms. Therefore two branes can only be dual if:

$$[D - (p_1 + 2)] = D - [D - (p_2 + 2)] \quad (6.203)$$

or in other words:

$$\boxed{p_1 + p_2 = D - 4} \quad (6.204)$$

So

- In 4-spacetime dimensions: Particles $p = 0$ are “dual” to particles $p = 0$.
- In 6-spacetime dimensions: Strings $p = 1$ are “dual” to strings $p = 1$.
- In 10-spacetime dimensions: Strings $p = 1$ are “dual” to 5-branes $p = 5$.

Note that the linking spheres of two such branes typically intersect in points.

6.13 Introduction to singularity theory

Roughly speaking, singularity theory is the study of maps of manifolds where the generic behavior we have described in the above sections fails. This happens when the rank of df jumps in some way.

This is an enormous subject and we will just touch a few points briefly.

6.13.1 Some motivations

1. *Saddle point techniques in field theory.* In field theory and statistical mechanics we are interested in integrals having the general form:

$$Z(g, \hbar) \equiv \int_{\mathbb{R}^n} \prod_{i=1}^n \frac{dx^i}{\sqrt{2\pi\hbar}} \exp \left[-\frac{1}{\hbar} S(x^1, \dots, x^n; g_\alpha) \right]$$

Here S is a real-valued function of the integration variables x^i and g_α , with suitable growth at infinity so the integral converges. We would like to know the dependence on parameters g_α (they could be coupling constants, temperature, magnetic fields, etc.) at least as $\hbar \rightarrow 0$. In this context, the parameters g_α are called *control parameters*. When applying the saddle-point method to such integrals one is interested in *critical points* of the function as a function of x^i (possibly analytically continued to complex x^i).

2. *Asymptotic behavior of special functions.* Many special functions of mathematics and mathematical physics are defined by integral representations depending on parameters. As just one example we consider the Airy function

$$\text{Ai}(\omega) := \int_{-\infty}^{+\infty} \frac{dx}{2\pi} e^{i\left(\frac{x^3}{3} + \omega x\right)} \quad (6.205)$$

Once again, ω now plays the role of a “control parameter.” It is important in many applications to understand the (somewhat subtle) asymptotic behavior of $\text{Ai}(\omega)$ as a function of ω for large $|\omega|$.

3. *Caustics in optics.* In geometrical optics the intensity of light at a point y , denoted $I(y)$ is related to an amplitude $A(y)$ by $I(y) = |A(y)|^2$ where

$$A(y) = \frac{1}{\lambda} \int_S \psi(x) e^{\frac{2\pi i}{\lambda} \Phi(x;y)} d^2x \quad (6.206)$$

where λ is the wavelength of the light, S is the source of the light, and $\psi(x)$ is the amplitude of the light at the source, $\Phi(x;y)$ is the “optical path length” from x to y , and we have taken the “eikonal approximation.” For much more information see the classic textbook: Born and Wolf, *Principles of Optics*. Once again, in the short-wavelength approximation we are interested in the critical points of $\Phi(x;y)$ in x as a function of y .

4. *Equilibrium of forces.* Suppose $V(x;c)$ is a potential energy function as a function of state variables x , depending on control parameters c . Then the equations for zero-force are that $V(x;c)$ has a critical point as a function of x .
5. *Thermodynamics and phase transitions.* Similarly, we could have a free energy function $F(\phi;c)$ for some fields or order parameters as a function of couplings or other external parameters c . A groundstate would be obtained by minimizing F as a function of the order parameters ϕ . Call the minima $\phi^{(\alpha)}(c)$ (there might be more

than one so we label the by α). In general, if the control parameters are changed by a small amount the groundstate order parameters $\phi^{(\alpha)}(c)$ will change by a small amount. However, there are some cases where a small change in control parameter leads to a large change in $\phi^{(\alpha)}(c)$. This is what happens in phase transitions.

6. *Gradient flow.* We could also consider a potential $U(x; c)$ for gradient flow in some variables x^i . This is the set of one-parameter flows satisfying the differential equation

$$\frac{dx^i}{dt} = -\frac{\partial U}{\partial x^i} \quad (6.207)$$

Again, we can ask how the qualitative behavior of the flows changes as the control parameters are changed.

Exercise *Gradient flow*

- a.) Show that under gradient flow (6.207) the potential function U is always nonincreasing with time along the flow.
- b.) For which initial conditions is it strictly decreasing?
- c.) Let $U(x)$ be a polynomial in one variable such that $U'(x)$ has n real roots. Describe the gradient flows.
-

6.13.2 Canonical forms of functions

Let us consider a function $f : \mathbb{R}^n \times \mathbb{R}^k \rightarrow \mathbb{R}$, where we separate the domain variables into two types $(x; c) \in \mathbb{R}^n \times \mathbb{R}^k$ called *state variables* and *control parameters*. That is we should be thinking of families of functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$, parametrized by $c \in \mathbb{R}^k$. Put differently, we are studying generic k -dimensional subspaces of $\text{Map}(\mathbb{R}^n, \mathbb{R})$ where, to get precise statements, we must put some conditions on the kinds of maps we wish to consider.

Now, we ask how much we can simplify our function using (suitable) changes of variables.

For simplicity, we will consider our function to be real analytic (i.e. it has a convergent Taylor series expansion) and we consider real analytic changes of variables, so we consider k -dimensional subspaces of $\text{Map}^{\text{analytic}}(\mathbb{R}^n, \mathbb{R})$. The theory can be extended to smooth functions.

First, recall what we know from the implicit function theorem: If

$$df(x_0; c) = \frac{\partial f}{\partial x^i} \Big|_{x_0} dx^i \neq 0 \quad (6.208)$$

then we can solve for x in terms of f . Put differently, we can find a change of variables

$$x^i \rightarrow y^i = a_j^i(x - x_0)^j + a_{jk}^i(x - x_0)^j(x - x_0)^k + \cdots \quad (6.209)$$

where we allow the $a_{j_1 j_2 \dots}^i$ to depend on c so that we can transform $f(x^i(y); c) = y^1$, valid in a neighborhood of x_0 .

Next, suppose x_c is a critical point, so $df(x_c; c) = 0$. Recall that the Hessian of f is the matrix of second derivatives at x_c :

$$f = \frac{1}{2}H_{ij}(x^i - x_c^i)(x^j - x_c^j) + a_{ijk}(x^i - x_c^i)(x^j - x_c^j)(x^k - x_c^k) + \cdots \quad (6.210)$$

We stress that the coordinates of the critical point x_c^i and the Taylor coefficients, including H_{ij} , are functions of c . The critical point is a *Morse critical point* if $\det H_{ij} \neq 0$, and then we have the important:

Morse Lemma: Near a Morse critical point, by a smooth change of variables we can find a change of variables $y^i(x; c)$ so that $x = x_c$ corresponds to $y^i = 0$ and, in some neighborhood of $y = 0$ we have:

$$f = \sum_{i=1}^n \lambda_i(c) y_i^2 \quad (6.211)$$

Proof...

Now, what happens if, as a function of control parameters $H_{ij}(c) = H_{ij}(x_c(c); c)$ becomes degenerate? Then we have the Thom Theorem:

Thom Theorem: Suppose at a value $c = c_*$, ℓ eigenvalues λ_i vanish. Say $\lambda_i(c_*) = 0$ for $1 \leq i \leq \ell$ and $\lambda_i(c_*) \neq 0$ for $i > \ell$. Then by a change of variables (6.209) we can bring f to the form:

$$f = f_{NM}(y^1(x; c), \dots, y^\ell(x; c)) + \sum_{i=\ell+1}^n \lambda_i(c) y_i^2 \quad (6.212)$$

where

1. For $\ell + 1 \leq i \leq n$, $\lambda_i(c)$ are nonzero in some neighborhood of c_*
2. In a neighborhood of $(x_*; c_*) \in \mathbb{R}^n \times \mathbb{R}^k$ the non-Morse function has the form

$$f_{NM}(y) = G(y) + P(y; c) \quad (6.213)$$

where $P(y; c)$, the perturbation, vanishes at c_* and $G(y)$ is a “canonical germ” or “catastrophe germ” for the singularity.

Part two is unfortunately rather vague. $G(y)$ is, roughly speaking, the “simplest” form for the non-Morse function at $c = c_*$. The perturbation is the simplest form for the non-Morse function we obtain by a generic perturbation of f in a k -dimensional subspace of $\text{Map}^{\text{analytic}}(\mathbb{R}^n, \mathbb{R})$ near $f(x; c_*)$. (To say it right we would need to talk about germs and jets. See the books of Arnold listed below for precise statements.)

Example: For example, suppose $n = 1$ and all the Taylor coefficients at $c = c_*$ below x^V vanish so that

$$f_{NM}(x; c_*) = a_V x^V + a_{V+1} x^{V+1} + \cdots \quad (6.214)$$

where $a_V \neq 0$. Then, by an analytic change of coordinates we can determine all the coefficients α_n in

$$x = \alpha_1 y + \alpha_2 y^2 + \cdots \quad (6.215)$$

so that

$$f_{NM}(x(y); c_*) = \pm y^V \quad (6.216)$$

Indeed, substitution of (6.215) into (6.214) gives an upper triangular system of equations on the α_n allowing us to remove the higher-order coefficients in (6.214) and bring it to the form (6.216). In the case of real variables, if V is even then we can only reduce a_V to its sign.

Now, if we perturb c away from c_* in the k -dimensional family of functions we will change

$$f_{NM}(y; c) \rightarrow \pm y^V + (\epsilon_0 + \epsilon_1 y + \epsilon_2 y^2 + \dots) \quad (6.217)$$

where the ϵ_n are functions of c which all vanish at $c = c_*$. Once again, by analytic redefinition of y we can remove all the terms above y^V . By a linear shift of y we can then remove the order y^{V-1} term, but then we are stuck. Putting these two steps together, our change of variables $x = \sum_{j=0}^{\infty} \alpha_j(c) y^j$ can be chosen so that for c near c_*

$$f_{NM}(x(y); c) = \pm y^V + \sum_{i=0}^{V-2} a_i(c) y^i \quad (6.218)$$

where $a_i(c_*) = 0$. Now, by further change of variables we *cannot*, in general remove the lower order terms in a family, with the exception of a_{V-1} which can be removed by a linear shift of y .

In this example $G(y) = \pm y^V$ and $P(y; c) = \sum_{i=1}^{V-2} a_i(c) y^i$. Shifting by $P(y; c)$ changes the qualitative behavior of $G(y)$, when $V > 2$.

Next, we need the notion of a *simple singularity* or an *elementary singularity*. If the Hessian $H(f)$ vanishes at $c = c_*$ then the Taylor series for f begins at order three:

$$f = \sum_{i,j,k} a_{ijk} x^i x^j x^k + \dots \quad (6.219)$$

we can ask if we can make an analytic change of coordinates

$$x^i = B_j^i y^j + B_k^i y^j y^k + \dots \quad (6.220)$$

to put the leading term into canonical form. In general there will be unremovable parameters. The leading coefficient a_{ijk} is a totally symmetric tensor in n -dimensions and therefore has

$$\left(\binom{n}{3} \right) = \frac{n(n+1)(n+2)}{3!} \quad (6.221)$$

independent terms. Since there are n^2 matrix elements, there will be free parameters when $\left(\binom{n}{3} \right) > n^2$, that is, when $n \geq 3$.

The singularity germs with no free parameters are called *elementary singularities*. The canonical germs and perturbations have been classified by Arnold:

Name	Germ	Perturbation
$A_k^\pm$	$\pm x^{k+1}$	$\sum_{i=0}^{k-1} a_i x^i$
$D_{2k}^\pm$	$\pm y^{2k-1} + x^2 y$	$\sum_{i=0}^{2k-3} a_i y^i + b_1 x + b_2 x^2$
$D_{2k+1}^\pm$	$\pm (y^{2k} + x^2 y)$	$\sum_{i=0}^{2k-2} a_i y^i + b_1 x + b_2 x^2$
$E_6^\pm$	$\pm (y^4 + x^3)$	$\sum_{i=0}^2 a_i y^i + \sum_{j=3}^5 b_j x y^{j-3}$
E_7	$x y^3 + x^3$	$\sum_{i=0}^4 a_i y^i + \sum_{j=5}^6 b_j x y^{j-5}$
E_8	$y^5 + x^3$	$\sum_{i=0}^3 a_i y^i + \sum_{j=4}^7 b_j x y^{j-4}$

1. In contrast to some tables we have kept the constant perturbation a_0 as one of the perturbations. It is often dropped because it can be absorbed in a constant shift of f , and it doesn't change many qualitative aspects of f . It *does* change qualitatively the behavior of some things, like roots.
2. $A_k^\pm$ are equivalent for k even.
3. The refinements of $\pm$ above disappear if we consider complex functions of complex variables.
4. The space of perturbations forms a vector space. This vector space is canonically isomorphic with

$$\mathbb{R}[x^i]/(dG) \quad (6.222)$$

where we divide the polynomial ring $\mathbb{R}[x^1, \dots, x^n]$ by the ideal generated by the functions $\partial_i G$, where G is the catastrophe germ. The perturbations given above clearly descend to elements of this vector space.

5. The relation to simple Lie groups is not an accident.
6. There is a rich extension of this theory to holomorphic functions of complex variables. Among its many applications, it has found use in string theory and supersymmetric field theory.

Exercise

Consider gradient flow in the (x, y) with a potential $U(x; c)$ given by the simple singularities with generic perturbation. (For $A_k^\pm$ add $+y^2$ to the potential).

Draw the flow lines between the critical points. ⁴³

⁴³ *Answer:* If you do it right, you should get the Dynkin diagrams of the corresponding simple Lie algebra.

6.13.3 Germs, Jets, and Unfoldings

Here we give a few formal definitions needed to make some of the above statements more precise.

First, one must put a topology on the space of analytic mappings $\mathbb{R}^n \rightarrow \mathbb{R}^m$. This can be done by using the Taylor series coefficients around $x = x_0 \in \mathbb{R}^n$. We define a subbasis for the topology by using the open balls $B(f; x_0; \epsilon)$ around a function f defined by saying it is the set of functions g whose Taylor series coefficients satisfy

$$\sum_{a=1}^m \sum_{n=0}^{\infty} \sum_{i_1, \dots, i_n} |f_{i_1 \dots i_n}^a - g_{i_1 \dots i_n}^a|^2 < \epsilon \quad (6.223)$$

Now, we define an equivalence relation $f \sim g$ if f can be brought to g by an analytic redefinition of (domain) coordinates. Thus, $x^2 \sim x^2 + 2x + 1$, and more generally $x^2 \sim x^2 + 2ax + a^2$ for any real number a .

Definition f is *stable* if there exists a neighborhood $\mathcal{U}$ of f so that for all $g \in \mathcal{U}$, g is equivalent to f .

Note:

1. Morse functions $\mathbb{R}^n \rightarrow \mathbb{R}$ are locally stable.
2. The class of Morse functions can be shown to be open and dense in the above topology.

Definition The *germ* of a function is an equivalence class of functions where $f \sim g$ if there exists an open set $U \subset \mathbb{R}^n$ on which $f|_U = g|_U$.

Let $\mathcal{G}_n$ be the set of germs of all functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ with $f(0) = 0$. It is a vector space, in fact, it is an algebra.

Definition:

- a.) A *k-parameter unfolding* of a germ $[f]$ is a germ of the form $[F]$ where

$$F : \mathbb{R}^n \times \mathbb{R}^k \rightarrow \mathbb{R} \quad (6.224)$$

$$F(x, 0) = f(x) \quad (6.225)$$

- b.) An unfolding is a *universal unfolding* if it is stable and minimal in k .

Theorem[Mather]: f has a universal unfolding iff the vector space

$$\mathcal{G}_n / \left(\frac{\partial f}{\partial x_i} \right) \quad (6.226)$$

is finite dimensional.

6.13.4 Some Examples

Let us consider some examples of unstable functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and their unfoldings.

Definition: The *bifurcation set*, or *discriminant locus* of $f(x; c)$ is the set of control parameters for which some critical points are non-Morse.

Example 1: Consider $f(x) = x^2$. This has a critical point at $x = 0$ and it is Morse. Note that it is stable in the following sense: If we make a perturbation to

$$f(x) = x^2 + a_1x + a_0 \quad (6.227)$$

then the critical point shifts, (to $x_c = -a_1/2$) but it is still a Morse critical point. By a real change of variables and a shift of $f(x)$ by a constant we can change coordinates to $\tilde{f} = y^2$. Thus, a generic shift of a quadratic remains diffeomorphic to a quadratic. This generalizes to n -dimensions: If $f(x) = \sum \lambda_i x_i^2$ and all the λ_i are nonzero then a generic perturbation of $f(x)$ can be brought to quadratic form by a diffeomorphism. Thus, Morse functions are stable.

Example 2: Now consider $f(x) = x^3$. Now if we make a perturbation

$$f(x; a) = x^3 + ax \quad (6.228)$$

where $a \neq 0$ the nature of the critical point depends on the control parameter a . If $a > 0$ there is no (real) critical point. If $a < 0$ there are two real critical points, and they are both Morse. These critical points are $\pm(-a/3)^{1/2}$ and as $a \nearrow 0^-$ they merge to a double root and then move off in complex conjugate pairs into the complex plane. This reflects a qualitative change in the graph of $f(x)$ as the control parameter a is changed. See Figure 45. The function $f(x) = x^3$ is unstable. After making a small perturbation we get functions which cannot be brought to the form $f(y) = y^3$ by any diffeomorphism, since a diffeomorphism will not change, for example, the number of roots.

Example 3: Now consider $f(x; a, b) = \frac{1}{4}x^4 + \frac{1}{2}ax^2 + bx$, where $(a, b) \in \mathbb{R}^2$ are control parameters. Then the equation for the critical point is the equation for the root of a cubic $p(x) = x^3 + ax + b = 0$, and we have analyzed above in Figure 33 how the roots behave as a function of (a, b) . The critical points are Morse if $3x_c^2 + a \neq 0$. That is, they fail to be Morse if

$$\begin{aligned} x_c^3 + ax_c + b &= 0 \\ 3x_c^2 + a &= 0 \end{aligned} \quad (6.229)$$

For *generic* a, b the discriminant $\Delta = 4a^3 + 27b^2 \neq 0$ and the critical points are all Morse. The bifurcation set is the locus in the (a, b) plane where $\Delta = 0$. When (a, b) move on a path crossing the bifurcation set, and the path crosses the set at a generic point two critical points merge and the qualitative nature of the function $f(x; a, b)$, as a function of x ,

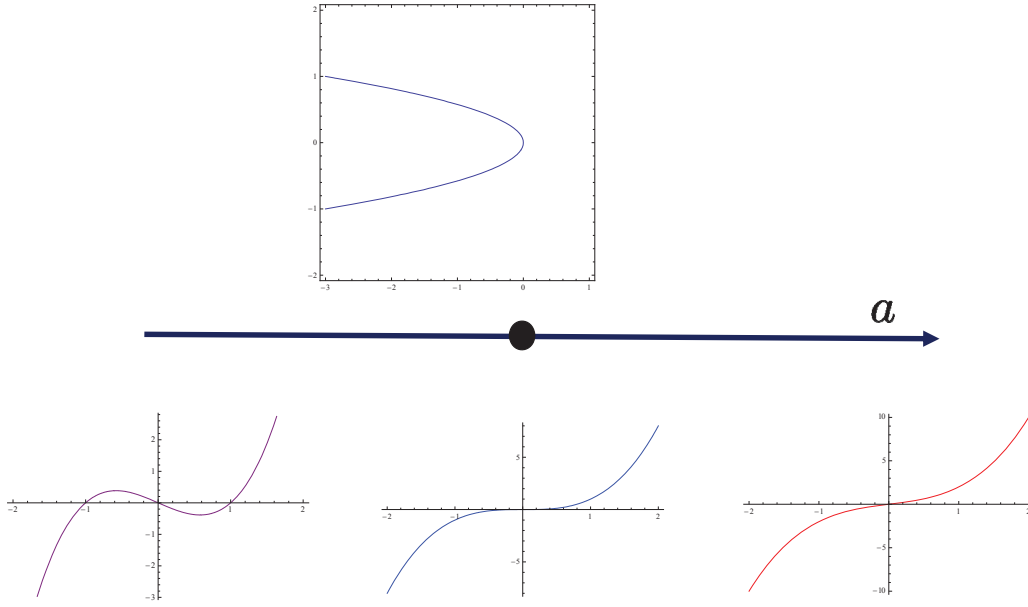


Figure 45: We plot the two critical points as a function of the control parameter a . Typical curves associated with a in the two regions separated by the separatrix $a = 0$ are shown below. $a = 0$ is also called the bifurcation set.

changes. See figures 46 and 47. At the nongeneric point, the tip of the cusp $\Delta = 0$ all three critical points merge to give $f(x) = x^4/4$.

A common application is in the theory of phase transitions in thermodynamics where $f(x; a, b)$ is a potential energy, x is a state variable and a, b are control parameters such as temperature and magnetization. If the control parameters evolve at a constant velocity then the state variable will jump suddenly, or discontinuously at the point p_5 in Figure 47. This is easily understood from the shape of the potential function as shown in Figure 48.

Remark: The sudden change of the state variable as a function of control variables in the above example (and in other similar examples) is the origin of the term “catastrophe theory.” In the early 1970’s it was proposed by the topologists René Thom and Christopher Zeeman that there could be wide-ranging applications of topology, specifically of results on the singularities of differentiable functions, to many topics very far removed from the traditional uses in mathematics (particularly differential equations), physics and engineering. See, for example, the book: Zeeman, E. C. (1977), *Catastrophe Theory: Selected*

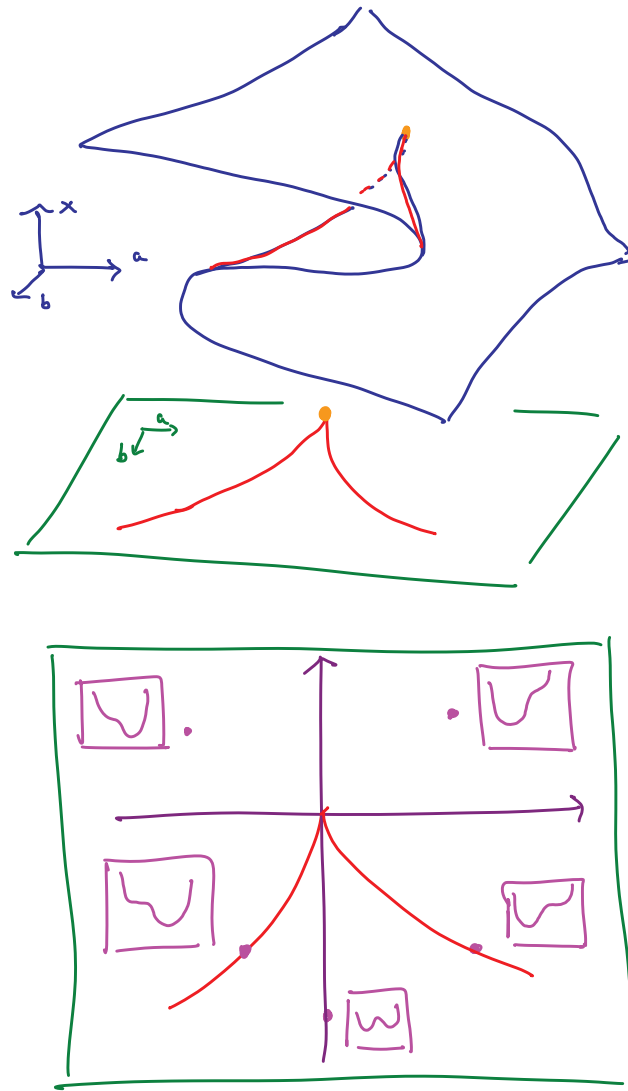


Figure 46:

Papers (1972-1977), Reading, MA: Addison-Wesley, where the term “catastrophe theory” was coined. Applications were proposed in economics, biology, social sciences, behavioral sciences, etc. It lead to a lot of attention in the popular press and generated a lot of controversy.⁴⁴

6.13.5 Maps between manifolds

TO BE WRITTEN

⁴⁴See, for just one example, M.W. Browne, “Experts Debate The Prediction Of Disasters; A Theory of Forecasting Events Prompts Bitter Scientific Debate,” New York Times, Nov. 19, 1977, p.47

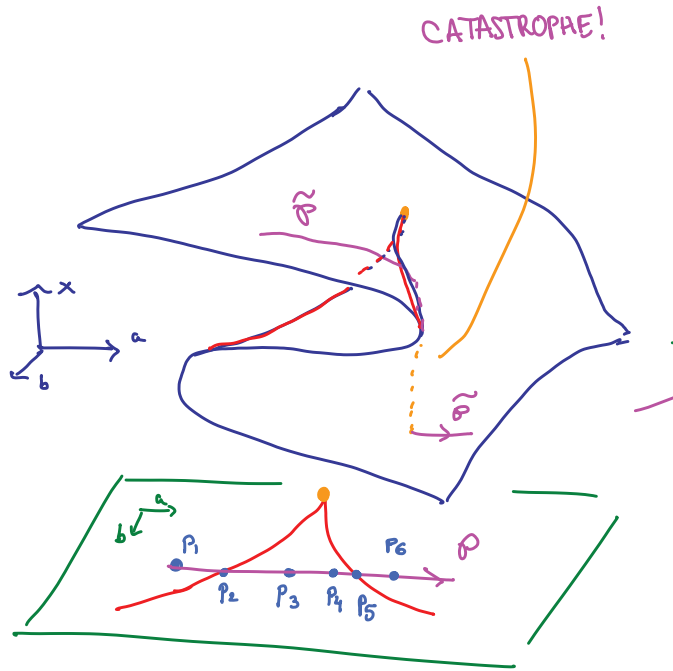


Figure 47: The control parameters move along the path φ . We attempt to find a lift of φ to describe a corresponding family of equilibria. Proceeding from p_1 to p_5 the state evolves smoothly, but then makes a “catastrophic” discontinuous jump as the control parameter is move to p_6 .

Theorem: A generic map $f : M_1 \rightarrow M_2$ for $\dim M_2 > 2\dim M_1$ is an embedding.

Whitney: Three stable types of $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$. Vision. Caustics.

6.13.6 Complex singularities

Holomorphic: $f : \mathbb{C}^{n+1} \rightarrow \mathbb{C}$ with $f(0) = 0$.

Suppose 0 is an isolated critical value.

Suppose moreover that the only critical point in $f^{-1}(0)$ is $(z^1, \dots, z^{n+1}) = 0$. Such a critical point is said to be an *isolated critical point*.

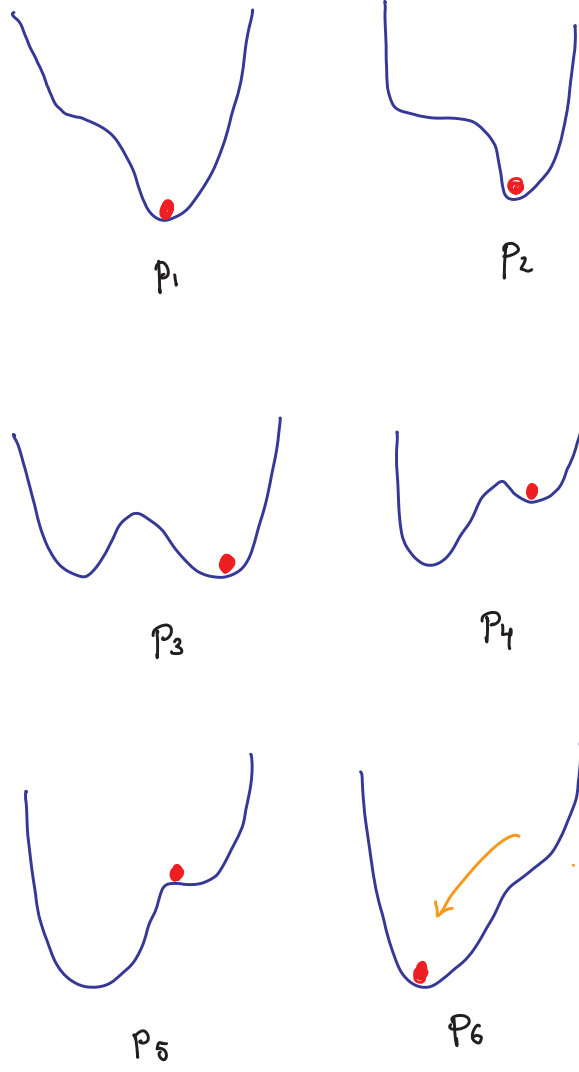


Figure 48: Profiles of the potential function for the path φ of control parameters.

$f^{-1}(t)$ is nonsingular real codimension two hypersurface of $\mathbb{C}^{n+1}$ for t in a sufficiently small punctured disk: $t \in B(0, \delta) - \{0\}$.

Now consider the sphere of radius-square $\epsilon > 0$ in $\mathbb{C}^{n+1}$:

$$S^{2n+1}(\epsilon) := \{z \mid |z^1|^2 + \cdots + |z^{n+1}|^2 = \epsilon\} \quad (6.230)$$

Then it is a theorem that there is an ϵ_0 such that for $0 < \epsilon < \epsilon_0$ the topology of the intersection $L(\epsilon) = f^{-1}(0) \cap S^{2n+1}(\epsilon)$ does not depend on ϵ , that is, $L(\epsilon)$ are homeomorphic for different ϵ . The topological space L is called the *link of the singularity*. Moreover, if 0 is an isolated singularity of f then the intersection is transverse and hence we can think of L as a compact real codimension two submanifold of S^{2n+1} .

Example: Consider $n = 1$. Then $L \subset S^3$ is a knot. For example, for $f(z) = z_1^3 + z_2^2$ the knot will be the trefoil.

Need to discuss Milnor algebra $\mathbb{C}[z^1, \dots, z^{n+1}]/(df)$.

6.13.7 Some sources

1. Arnold, *Catastrophe Theory*, Springer-Verlag. Popular level book.
2. J.W. Auer, "Mathematical Preliminaries to Elementary Catastrophe Theory," *Mathematics Magazine*, vol. 53, pp. 13-20. <http://www.jstor.org/stable/2690024>. Short undergraduate-level article.
3. R. Gilmore, *Catastrophe Theory for Scientists and Engineers*, Dover 1993
4. R. Gilmore, "Catastrophe Theory," *Encyclopedia of Applied Physics*, Vol. 3, 1992
5. V.I. Arnold, S.M. Gusein-Zade, and A.N. Varchenko, *Singularities of Differentiable Maps*, vol. 1 +2
6. V.I. Arnold, V.V. Goryunov, O.V. Lyashko, V.A. Vasil'ev, *Singularity Theory*, vol. 1 +2
7. J.W. Bruce and P.J. Giblin, *Curves and Singularities*, Cambridge, 1992
8. Saunders, *An introduction to catastrophe Theory*, Cambridge, 1980

For the holomorphic theory see:

1. V.I. Arnold, V.V. Goryunov, O.V. Lyashko, V.A. Vasil'ev, *Singularity Theory*, vol. 1 +2
2. J.W. Bruce and P.J. Giblin, *Curves and Singularities*, Cambridge, 1992
3. Dimca, *Singularities of Hypersurfaces*
4. J. Milnor, ...

6.14 Digression: Classification of manifolds

What can we say about the classification of manifolds?

A natural way to organize the problem is by dimension. The case of dimension one is easy and was discussed above. There is only one compact one-manifold without boundary, namely, the circle.

Manifolds of dimension two are more complicated. Of course we can take direct products to get $S^1 \times S^1 = T^2$, but of course there are many more examples, such as genus g surfaces and their nonorientable counterparts. In the case of $n = 2$ we have a beautiful

classification. We discuss it in Section §10. They are classified by: Orientability (or not), number of boundary components, and Euler character.

The situation is completely different in dimensions three and larger where the problem is much harder. Of course, we could take products like $S^1 \times \Sigma$, where Σ is a two-dimensional manifold. But many more possibilities arise. An important fact is that for a surface Σ the group of diffeomorphisms $\text{Diff}(\Sigma)$ has a natural topology and it is disconnected (except when $\Sigma = S^2$). For example $\pi_0(\text{Diff}(T^2)) \cong PGL(2, \mathbb{Z})$. For higher genus surfaces the group is known as the Teichmüller group and is rather nontrivial.

Now, consider an arbitrary knot $K \subset S^3$. Its tubular neighborhood is diffeomorphic to $S^1 \times D^2$ and has boundary $S^1 \times S^1$. Thus, we could cut out the neighborhood and glue it back in with a general diffeomorphism of the torus. Similarly, we could take two genus g handlebodies, (analogous of the solid torus with boundary a genus g surface Σ). Take a diffeomorphism $\phi \in \text{Diff}(\Sigma)$ and glue the handlebodies together. Clearly, there is a large universe of complicated three manifolds.⁴⁵

Do you think it gets better, or worse, in higher dimensions?

6.14.1 Three categories of manifolds

In discussing this topic an essential point is that there are different kinds of manifolds depending on what conditions we put on the gluing functions

$$\phi_{\alpha\beta} : \phi_\alpha(\mathcal{U}_\alpha \cap \mathcal{U}_\beta) \rightarrow \phi_\beta(\mathcal{U}_\alpha \cap \mathcal{U}_\beta) \quad (6.231)$$

used to construct a manifold from an atlas.

These lead to three categories that are commonly discussed:

TOP: The $\phi_{\alpha\beta}$ are homeomorphisms, and morphisms of manifolds are continuous maps in local coordinate charts.

PL: The $\phi_{\alpha\beta}$ and morphisms of manifolds are “piecewise linear on suitably fine triangulations.”

DIFF: The $\phi_{\alpha\beta}$ are diffeomorphisms, and morphisms of manifolds are smooth maps in local coordinate charts.

Remarks:

1. In physics we generally work with differentiable manifolds and maps of manifolds, although it is important to allow singularities.
2. The category **PL** is not often used in physics (except implicitly in statistical mechanics and lattice field theory). It is rather technical so we won’t give precise definitions. Roughly speaking, it means the manifold admits a triangulation. The gluing functions are piecewise affine-linear transformations.

⁴⁵A decomposition of a closed 3-manifold of the above type is known as a *Heegaard splitting*. By a theorem of Moise, every 3-manifold can be triangulated. Take the 1-skeleton and thicken it to get a handlebody. This shows that every three-manifold admits a Heegaard decomposition. The trouble is, there are a lot of nontrivial diffeomorphisms of a surface to itself, and it can be hard to recognize two equivalent 3-folds constructed from different Heegaard splittings.

It should be plainly evident that every smooth manifold also defines a topological manifold, since a diffeomorphism is *a fortiori* also a homeomorphism: We just forget the differentiability. Similarly, a PL manifold also defines a topological manifold. Thus **DIFF** $\subset$ **TOP** and **PL** $\subset$ **TOP**.

It is nontrivial, but true, that a smooth manifold is uniquely a PL manifold. This is a theorem of Whitehead. Thus we have

$$\mathbf{DIFF} \subset \mathbf{PL} \subset \mathbf{TOP} \quad (6.232)$$

In dimensions 1, 2, 3 the three categories are “the same:” Every topological manifold admits a unique PL or smooth structure. (Unique up to PL or smooth equivalence.) However, in dimensions four and above these three categories are very different. We will list a few facts.

Let us start with a manifold M in **TOP**. We should ask two fundamental questions:

1. Can we put M in **PL**? That is, roughly speaking, can we triangulate M ?
2. If M can be triangulated, can it be triangulated in a unique way?

A theorem of Moise states that the answer to both question is “yes” if $\dim M \leq 3$. However, in higher dimensions the answer can be “no.” Kirby and Siebenmann showed that for compact manifolds without boundary of dimension $\dim M \geq 5$ there is a single topological invariant

$$\kappa(M) \in H^4(M; \mathbb{Z}_2) \quad (6.233)$$

(we are not explaining the precise definition of $\kappa(M)$ here) which “measures” the obstruction to a PL structure. That is, if M is compact without boundary and $\dim M \geq 5$ then when $\kappa(M) = 0$ a PL structure exists and if it is nonzero there is no PL structure, and hence no smooth structure.

Furthermore, if $\dim M \geq 5$ and M is PL, then a PL map $f : M \rightarrow M$ has an associated invariant $\theta(f) \in H^3(M; \mathbb{Z}_2)$. If $\theta(f)$ is not zero the PL structures are inequivalent. Moreover, for every element in $H^3(M; \mathbb{Z}_2)$ there is such a map. Hence, for $\dim M \geq 5$ PL structures on a given topological manifold are classified.

If $\dim M = 4$ and $\kappa(M)$ is nonzero then there is no PL structure but there can be other obstructions to putting a smooth structure on M . See below.

Next, we ask the analogous two fundamental questions in going from **PL** to **DIFF**: If M in **PL** then

1. Can we put a smooth structure on M ?
2. If so, is it unique?

Regarding PL structures we have:

Theorem Every PL manifold of dimension $n \leq 7$ has a compatible smooth structure. Moreover this structure is unique for $n < 7$.

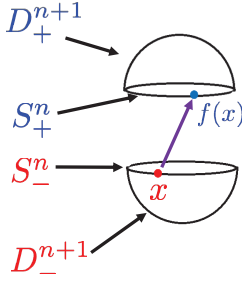


Figure 49: How to make an exotic $(n + 1)$ -dimensional sphere: Glue together two $(n + 1)$ -disks along their common S^n boundary by using a diffeomorphism $f : S^n \rightarrow S^n$ which is not continuously connected to the identity.

In $n = 7$ dimensions something dramatic happens: There exist *exotic spheres*. That is, there are manifolds in **TOP** which are homeomorphic to the standard S^7 but are not diffeomorphic to it! This was a great shock to the mathematical community when J. Milnor discovered it in 1956. ⁴⁶

The essential fact is that the *group of diffeomorphisms* $\text{Diff}(S^n)$, itself a topological group, can have many connected components. To make an exotic $(n + 1)$ -sphere you glue two $(n + 1)$ -disks together using a diffeomorphism f from a nontrivial component of $\text{Diff}(S^n)$ as in Figure 49. A nontrivial result shows that every exotic sphere is of this form. It turns out that $\text{Diff}(S^6)$ has 28 connected components.

Actually, the exotic 7-spheres are really not that exotic. They have very concrete models. Milnor's original discussion involved perfectly natural S^3 fibrations over S^4 . One can also relate them to complex singularities:

Theorem Consider the set of functions $f_\ell : \mathbb{C}^5 \rightarrow \mathbb{C}$ enumerated by $1 \leq \ell \leq 28$ and defined by

$$f_\ell(z) = z_1^2 + z_2^2 + z_3^2 + z_4^3 + z_5^{6\ell-1} \quad (6.234)$$

Then $z = 0$ is an isolated singularity and the link L is homeomorphic to the ordinary S^7 and differentiable to the ℓ^{th} Milnor exotic sphere.

6.14.2 Four dimensions

The situation in four dimensions is somewhat special and not completely understood, although some of the greatest advances in 20th century mathematics have gone a long way towards solving the problem.

The first (very easy) result of Markov might seem a little discouraging:

Theorem Any finitely generated group G is the fundamental group of some four-manifold.

⁴⁶See J. Milnor, Ann. Math. **64**(1956)399; Kervaire and Milnor, Ann. Math. **77**(1963)505

Proof: Suppose the group G has presentation:

$$G \cong \langle g_1, \dots, g_n | R_1, \dots, R_m \rangle \quad (6.235)$$

We aim to produce a four-manifold M_4 with fundamental group G . First, consider the free group on one generator $\langle g \rangle \cong \mathbb{Z}$. A good manifold that has this as a fundamental group is $X_4 = S^1 \times S^3$. Now let us consider $\tilde{M}_4 := X_4 \# \dots \# X_4$. Then

$$\pi_1(\tilde{M}_4) \cong \langle g_1, \dots, g_n \rangle \quad (6.236)$$

is the free group with generators g_i corresponding to the simple loops around the S^1 factor in each summand (extended to some common basepoint). Now, each relation R_α is a word in the g_i and thus can be represented by some closed based loop $\ell_\alpha \subset \tilde{M}_4$. We can take the ℓ_α to be nonintersecting, by simple codimension arguments. Now, take a tubular neighborhood $N(\ell_\alpha)$ of ℓ_α . By our discussion above of the local picture of submanifolds it is diffeomorphic to $N(\ell_\alpha) \cong S^1 \times D^3$, where D^3 is the 3-dimensional ball. The boundary is thus $\partial N(\ell_\alpha) \cong S^1 \times S^2$. This is also the boundary of $D^2 \times S^2$. So, glue in a copy of $D^2 \times S^2$ along the boundary of $N(\ell_\alpha)$. This procedure is known as *urgery*. Now the loop S^1 in $S^1 \times D^3$ (which was representing the word R_α) becomes contractible! Thus it is a relation on the generators g_i in the new manifold. We can choose the tubular neighborhoods around the different loops ℓ_α to be nonintersecting, and hence we can perform surgeries on each of these loops without interference. If we do this for all the loops we produce our manifold M_4 . By the Seifert-van Kampen theorem (see Section 12.2 below) it follows that the fundamental group of M_4 is exactly G . ♠

Since finitely presented groups cannot be classified it follows that four-manifolds cannot be classified, even up to homotopy type. We therefore retreat and ask about the classification of *simply connected* four-manifolds.

Theorem[Whitehead, Milnor] Two simply connected four-manifolds are homotopy equivalent iff their intersection forms are equivalent $Q(X_1) \cong Q(X_2)$.

The question of topological equivalence is much more difficult. The main result is a complete classification of simply connected closed compact topological four-manifolds. It is due to Michael Freedman:

Theorem[Freedman]. A complete invariant for a closed simply connected ⁴⁷ topological four manifold M is the pair $(Q(M), \kappa(M))$ where ⁴⁸

1. $Q(M)$ is the intersection form $\bar{H}_2(M; \mathbb{Z}) \otimes \bar{H}_2(M; \mathbb{Z}) \rightarrow \mathbb{Z}$
2. $\kappa(M) \in H^4(M; \mathbb{Z}_2) \cong \mathbb{Z}_2$ is the Kirby-Siebenmann invariant.

⁴⁷Any simply connected manifold is orientable. Changing the orientation on M changes the overall sign of the intersection form.

⁴⁸Here $\bar{H}_2(M; \mathbb{Z}) = H_2(M; \mathbb{Z}) / \text{Tors}(H_2(M; \mathbb{Z}))$.

Moreover,

1. For every symmetric integral form Q of determinant 1 there is a topological manifold M such that $Q(M) = Q$.
2. If $Q(M)$ is such that (x, x) is even for every $x \in H_2(M; \mathbb{Z})$ then $\kappa(M)$ is not independent and $\kappa(M) = \frac{1}{8}\text{Sig}(Q(M)) \bmod 2$.
3. If Q is such that there is an $x \in H_2(M; \mathbb{Z})$ with (x, x) an odd integer then for either of the two possible values of κ there is manifold with $Q = Q(M)$ and $\kappa = \kappa(M)$.

To get a sense of the breathtaking implications of this consider the following two examples:

Example 1: If M is homotopic to the four-dimensional sphere then $Q(M) = 0$. Therefore M is homeomorphic to S^4 . This solves the four-dimensional topological Poincaré conjecture.

Example 2: Suppose $M = \mathbb{CP}^2$. Let $a_1X^1 + a_2X^2 + a_3X^3$ be any nonzero linear form. Then the equation

$$a_1X^1 + a_2X^2 + a_3X^3 = 0 \tag{6.237}$$

in homogeneous coordinates makes sense and defines a one-complex-dimensional submanifold in $\mathbb{CP}^2$. In fact, by a linear transformation, we might as well say it is $X^3 = 0$. But clearly the set of elements

$$\Sigma := \{[X^1 : X^2 : 0]\} \subset \mathbb{CP}^2 \tag{6.238}$$

is a copy of $\mathbb{CP}^1 \cong S^2$. One can show that $[\Sigma]$ generates $H_2(\mathbb{CP}^2; \mathbb{Z}) \cong \mathbb{Z}$. Therefore, the only relevant intersection number is $\iota(\Sigma, \Sigma)$. We can compute it by perturbing Σ to $\tilde{\Sigma}$ so that Σ and $\tilde{\Sigma}$ have a transverse intersection. So we perturb the linear form to define:

$$\tilde{\Sigma} := \{[X^1 : X^2 : X^3] | \epsilon_1X^1 + \epsilon_2X^2 + X^3 = 0\} \tag{6.239}$$

where at least one of ϵ_i is nonzero. The intersection is the set of $[X^1 : X^2 : X^3]$ such that $X^3 = 0$ and

$$\epsilon_1X^1 + \epsilon_2X^2 = 0 \tag{6.240}$$

If, say, $\epsilon_1 \neq 0$ then any point in the intersection must be of the form $[-\frac{\epsilon_2}{\epsilon_1}X^2 : X^2 : 0]$. Clearly X^2 cannot be zero. But then we can scale it to 1 so the intersection $\Sigma \cap \tilde{\Sigma} = \{[-\frac{\epsilon_2}{\epsilon_1} : 1 : 0]\}$ consists of a single point. (It is easy to show that there are canonical orientations so the intersection number is naturally +1 and not -1.) Thus, Q is just multiplication $\mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}$. Since $\mathbb{CP}^2$ is a smooth four-manifold we know that $\kappa(\mathbb{CP}^2) = 0$. Now, by Freedman's theorem we know that there must exist another four manifold M' - we could call it a "fake $\mathbb{CP}^2$ " - with $Q = 1$ and $\kappa = 1$, which is not smoothable.

The methods used by Freedman were, to quote Milnor, "wildly non-differentiable."

Indeed, the classification of four-manifolds in **PL** and **DIF** is at present quite mysterious.

The main source of knowledge is an important set of theorems of S. Donaldson which use properties of the space of solutions to Yang-Mills equations:

$$D^\mu F_{\mu\nu} = 0 \quad (6.241)$$

(Where A_μ is a connection on a principal G bundle over M . See below.) One can formulate Yang-Mills theory on a four-dimensional *smooth* manifold. Since the Yang-Mills equations are differentiable equations, the smooth structure is used here. Indeed, the construction uses a *Riemannian metric* $g_{\mu\nu}dx^\mu \otimes dx^\nu$! The solutions of the equations with minimal Yang-Mills action:

$$S[A] = \int_M \text{Tr}(F^{\mu\nu} F_{\mu\nu}) \text{vol}(g) \quad (6.242)$$

come in families, or “moduli spaces.” The key observation is that the minimal action gauge fields satisfy a first order equation because:

$$S[A] = \int_M \text{Tr}(F^{\mu\nu} F_{\mu\nu}) \text{vol}(g) = \frac{1}{2} \int_M \text{Tr}(F \pm *F)^2 \mp \int_M \text{Tr}(F \wedge F) \quad (6.243)$$

Now $\int_M \text{Tr}(F \wedge F)$ is a topological invariant (of the gauge bundle) and does not change under continuous deformation of gauge field. Since $S[A]$ is clearly nonnegative we should choose the sign so that $\mp \int_M \text{Tr}(F \wedge F) = |\int_M \text{Tr}(F \wedge F)|$ and then the minimum is achieved when $F \pm *F = 0$. This is a first order differential equation for A and hence much more tractable than the original YM equations. The equations and action all depend on the Riemannian metric. However, (with the exception of a well-understood class of four-manifolds) the metric dependence does not affect the topology of the moduli space of instantons. Topological invariants of these moduli spaces provide smooth invariants of the original four-manifold M . The use of the Yang-Mills equations here is a notable example of the influence of physics in mathematics.

One important consequence of Donaldson’s investigation of these moduli spaces is the following theorem: ⁴⁹

Theorem: If the intersection form of a simply connected manifold is positive definite or negative definite then it is diagonalizable (over the integers!).

On the other hand, from the theory of integral quadratic forms we know that there are lots of unimodular symmetric forms which cannot be diagonalized. Using Freedman’s theorem we conclude that these manifolds admit no smooth structure!

⁴⁹The idea of the proof is very simple: For a simple choice of bundle on which the Yang-Mills connection is defined the moduli space is a five dimensional manifold with “ends.” At one end it is a copy of M because all the fieldstrength becomes concentrated at a single point $p \in M$, and this can happen at any point p . On the other hand, there will be a finite set of degenerate connections and in the neighborhood of these connections the moduli space looks like a cone over $\mathbb{CP}^2$ or its complex conjugate. Then one notes that the intersection form is a cobordism invariant.

The first example is the famous Cartan matrix of the Lie algebra E_8 . This produces the famous “ E_8 manifold.” In one basis it is

$$\begin{bmatrix} 2 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 \end{bmatrix} \quad (6.244)$$

Donaldson’s theorem gives some criteria for deciding when certain topological four-manifolds admit no smooth structure. It is still not a completely solved problem to say when a topological four-manifold does admit a smooth structure. This is the subject of something called the “11/8 conjecture.” The best progress on this problem used something called Seiberg-Witten invariants, a smooth manifold invariant coming directly out of physics.

Moreover, even when a topological four-manifold admits a smooth structure, it might admit many inequivalent smooth structures. That is, there can be two smooth four-manifolds M_1, M_2 which are homeomorphic but are not diffeomorphic. In fact, a given topological four-manifold can admit countably infinitely many different inequivalent smooth structures. At present there is not even a proposed classification scheme for smooth four-manifolds. The best progress has come from using Seiberg-Witten invariants. See the article of Fintushel and Stern referenced below for the state of the art.

One can show that one consequence of Donaldson’s theorems is that there is a continuum of inequivalent “fake $\mathbb{R}^4$ ’s.” These have compact sets which cannot be surrounded by diffeomorphically embedded 3-spheres. By contrast, all other $\mathbb{R}^n$ for $n \neq 4$ have a unique differentiable structure, except for $n = 4$.

It is quite striking that the smooth invariants have made use of Yang-Mills equations. In 1988 Witten showed that the Donaldson invariants could be identified with correlation functions of certain operators in a four-dimensional Yang-Mills-Higgs theory with extended supersymmetry. In the process he invented the very influential idea of topological field theory. In 1994, using physical insights provided by Seiberg and Witten about the structure of the ground states in such YMH theories, Witten introduced a new set of topological invariants, known as *Seiberg-Witten invariants* which are much easier to compute than the Donaldson invariants and can be used to prove many of the earlier theorems of Donaldson. Indeed, it has been shown that the Seiberg-Witten invariants contain the same amount of information as Donaldson invariants.

6.14.3 The Generalized Poincaré conjecture

The Poincaré conjecture asks whether an n -dimensional manifold M which has the homotopy groups of the n -dimensional sphere S^n is in fact equivalent to S^n . The notion of “equivalent” depends on which category we are in, **TOP**, **PL**, or **DIFF**, and the answer is different in each case.

After a great deal of work we know:

Theorem The topological Poincaré conjecture is true in all dimensions.

The last case to succumb was $n = 3$. See ⁵⁰

Once again, one of the key elements came from physics. D. Friedan introduced a differential equation associated with the renormalization flow of nonlinear sigma models:

$$\frac{d}{d\tau}g_{\mu\nu} = -R_{\mu\nu} \quad (6.245)$$

known as Ricci flow. Hamilton realized that Ricci flow held the key to understanding the Poincaré conjecture and Perelman was able to analyze it in sufficient detail to prove the conjecture for $n = 3$.

Theorem The PL Poincaré conjecture is true in all dimensions except possibly 4.

The smooth Poincaré conjecture is much more complicated. In fact, it is not true in dimension $n = 7$ as already described above.

Let $\mathcal{S}_n$ be the set of oriented diffeomorphism classes of n -manifolds with the homotopy type of S^n . This set has an abelian addition operation under connected sum and an identity given by the standard S^n . That is, it is an abelian monoid. It is not known whether $\mathcal{S}_n$ is finite or infinite for $n = 4$, but for all other dimensions it is known to be finite. Therefore, except possibly in $n = 4$ it is an abelian group. Although much is known, the full structure of this group is not yet known. Except for $n = 4$, the monoid is known for low dimensions:

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$\mathcal{S}_n$	1	1	1	?	1	1	28	2	(2, 2, 2)	6	992	1	3	2	(2, 8128)	2	(2, 8)	(2, 8)

In this table (taken from Milnor’s article cited below) in line two k stands for the abelian group $\mathbb{Z}/k\mathbb{Z}$ and (k_1, k_2) stands for $\mathbb{Z}/k_1\mathbb{Z} \oplus \mathbb{Z}/k_2\mathbb{Z}$, and so on.

The smooth Poincaré conjecture is true iff $\mathcal{S}_n$ consists of a single element. This is known to be true for $n = 1, 2, 3, 5, 6, 12, 61$ but it is unknown if it is true for any other values of n .

6.14.4 Sources

1. D. Freed and K. Uhlenbeck, *Instantons and Four-Manifolds*
2. J. Milnor, “Differential Topology Forty-six Years Later,” Notices Amer. Math. Soc., Vol. 58, p. 804

⁵⁰J. Morgan and G. Tian, *Ricci Flow and the Poincaré Conjecture*, Clay Math Monographs 3, AMS, 2007; J. Lott and ???

3. C. Nash, *Differential Topology and Quantum Field Theory*.
4. Wikipedia articles: 4-manifold, Hauptvermutung, Piecewise linear manifold, Generalized Poincaré Conjecture
5. A. Scorpan, *The Wild World of Four-Manifolds*
6. Fintushel and Stern, “Six lectures on four manifolds.”

For explanations of the physics approach to Donaldson theory and Seiberg-Witten invariants see:

1. J. Labastida and M. Marino, *Topological Quantum Field Theory and Four Manifolds*
2. G. Moore lectures at <http://scgp.stonybrook.edu/archives/1964>.

7. Transformation Groups, Group Actions, and Orbits

7.1 Definitions and the stabilizer-orbit theorem

Let X be any set (possibly infinite). Recall the definition from chapter 1:

Definition 7.1: A *permutation* of X is a 1-1 and onto mapping $X \rightarrow X$. The set S_X of all permutations forms a group under composition.

Definition 7.2a: A *transformation group* on X is a subgroup of S_X .

This is an important notion so let’s put it another way:

Definition 7.2b:

A *G-action* on a set X is a map $\phi : G \times X \rightarrow X$ compatible with the group multiplication law as follows:

A *left-action* satisfies:

$$\phi(g_1, \phi(g_2, x)) = \phi(g_1 g_2, x) \quad (7.1)$$

A *right-action* satisfies

$$\phi(g_1, \phi(g_2, x)) = \phi(g_2 g_1, x) \quad (7.2)$$

In addition in both cases we require that

$$\phi(1_G, x) = x \quad (7.3)$$

for all $x \in X$.

A set X equipped with a (left or right) G -action is said to be a G -set.

Remarks

1. If ϕ is a left-action then it is natural to write $g \cdot x$ for $\phi(g, x)$. In that case we have

$$g_1 \cdot (g_2 \cdot x) = (g_1 g_2) \cdot x. \quad (7.4)$$

Similarly, if ϕ is a right-action then it is better to use the notation $\phi(g, x) = x \cdot g$ so that

$$(x \cdot g_2) \cdot g_1 = x \cdot (g_2 g_1). \quad (7.5)$$

2. If ϕ is a left-action then $\tilde{\phi}(g, x) := \phi(g^{-1}, x)$ is a right-action, and vice versa. Thus there is no essential difference between a left- and right-action. However, in computations with nonabelian groups it is extremely important to be consistent and careful about which choice one makes.
3. A given set X can admit more than one action by the same group G . In that case, to avoid confusion one should take care to distinguish the G -actions unless a particular one is understood. For example one could write $\phi_g(x) = \phi(g, x)$ and speak of ϕ_g , rather than write $g \cdot x$.

There is some important terminology one should master when working with G -actions:

Definitions:

1. A group action is *effective* or *faithful* if for any $g \neq 1$ there is *some* x such that $g \cdot x \neq x$. Equivalently, the only $g \in G$ such that ϕ_g is the identity transformation is $g = 1_G$.
2. A group action is *transitive* if for any pair $x, y \in X$ there is some g with $y = g \cdot x$.
3. A point $x \in X$ is a *fixed point* of G if there exists an element $g \in G$ with $g \neq 1$ such that $g \cdot x = x$.
4. Given a point $x \in X$ the set of group elements:

$$G^x := \{g \in G : g \cdot x = x\} \quad (7.6)$$

is called the *isotropy group at x* . It is also called the *stabilizer group* of x . (The reader should show that $G^x \subset G$ is in fact a subgroup.)

5. Given a group element $g \in G$ the *fixed point set* of G is the set

$$\text{Fix}(g) := \{x \in X | g \cdot x = x\} \quad (7.7)$$

The fixed point set of g is often denoted by X^g .

6. A group action is *free* if for any $g \neq 1$ then for *every* x , we have $g \cdot x \neq x$. That is, for every x the stabilizer group G^x is the trivial subgroup $\{1_G\}$. Equivalently, for every $g \neq 1$ the set $\text{Fix}(g)$ is the empty set.
7. The *orbit of G through a point x* is the set of points $y \in X$ which can be reached by the action of G :

$$O_G(x) = \{y : \exists g \text{ such that } y = g \cdot x\} \quad (7.8)$$

Remarks:

1. If we have a G -action on X then we can define an equivalence relation on X by defining $x \sim y$ if there is a $g \in G$ such that $y = g \cdot x$. (Check this is an equivalence relation!) The orbits of G are then exactly the equivalence classes of under this equivalence relation.
2. The group action restricts to a transitive group action on any orbit.
3. If x, y are in the same orbit then the isotropy groups G^x and G^y are conjugate subgroups in G . Therefore, to a given orbit, we can assign a definite *conjugacy class* of subgroups.

Point 3 above motivates the

Definition If G acts on X a *stratum* is a set of G -orbits such that the conjugacy class of the stabilizer groups is the same. The set of strata is sometimes denoted $X // G$.

Exercise

Suppose X is a G -set.

- a.) Show that the subset H of elements which act ineffectively, i.e. the set of $h \in G$ such that $\phi(h, x) = x$ for all $x \in X$ is a normal subgroup of G .
 - b.) Show that G/H acts effectively on X .
-

Exercise

Let G act on a set X .

- a.) Show that the stabilizer group at x , denoted G^x above, is in fact, a subgroup of G .
 - b.) Show that the G action is free iff the stabilizer group at every $x \in X$ is the trivial subgroup $\{1_G\}$.
 - c.) Suppose that $y = g \cdot x$. Show that G^y and G^x are conjugate subgroups in G .
-

Exercise

- a.) Show that whenever G acts on a set X one can canonically define a groupoid: The objects are the points $x \in X$. The morphisms are pairs (g, x) , to be thought of as arrows $x \xrightarrow{g} g \cdot x$. Thus, $X_0 = X$ and $X_1 = G \times X$.

- b.) What is the automorphism group of an object $x \in X$.

This groupoid is commonly denoted as $X // G$.

7.1.1 The stabilizer-orbit theorem

There is a beautiful relation between orbits and isotropy groups:

Theorem [Stabilizer-Orbit Theorem]: Each left-coset of G^x in G is in 1-1 correspondence with the points in the G -orbit of x :

$$\psi : \text{Orb}_G(x) \rightarrow G/G^x \quad (7.9)$$

for a 1 – 1 map ψ .

Proof: Suppose y is in a G -orbit of x . Then $\exists g$ such that $y = g \cdot x$. Define $\psi(y) \equiv g \cdot G^x$. You need to check that ψ is actually well-defined.

$$y = g' \cdot x \quad \rightarrow \quad \exists h \in G^x \quad g' = g \cdot h \quad \rightarrow \quad g'G^x = ghG^x = gG^x \quad (7.10)$$

Conversely, given a coset $g \cdot G^x$ we may define

$$\psi^{-1}(gG^x) \equiv g \cdot x \quad (7.11)$$

Again, we must check that this is well-defined. Since it inverts ψ , ψ is 1-1. ♠

Corollary: If G acts transitively on a space X then there is a 1 – 1 correspondence between X and the set of cosets of H in G where H is the isotropy group of any point $x \in X$. That is, at least as sets: $X = G/H$. The isotropy groups for points in G/H are the conjugate subgroups of H in G .

Remark: Sets of the type G/H are called *homogeneous spaces*. This theorem is the beginning of an important connection between the *algebraic* notions of subgroups and cosets to the *geometric* notions of orbits and fixed points. Below we will show that if G, H are topological groups then, in some cases, G/H are beautifully symmetric topological spaces, and if G, H are Lie groups then, in some cases, G/H are beautifully symmetric manifolds.

Exercise The Lemma that is not Burnside's

Suppose a finite group G acts on a finite set X as a transformation group. A common notation for the set of points fixed by g is X^g . Show that the number of distinct orbits is the averaged number of fixed points:

$$|\{\text{orbits}\}| = \frac{1}{|G|} \sum_g |X^g| \quad (7.12)$$

For the answer see. ⁵¹

⁵¹ Answer: Write

$$\sum_{g \in G} |X^g| = |\{(x, g) | g \cdot x = x\}| = \sum_{x \in X} |G^x| \quad (7.13)$$

Exercise *Jordan's theorem*

Suppose G is finite and acts transitively on a finite set X with more than one point. Show that there is an element $g \in G$ with no fixed points on X .⁵²

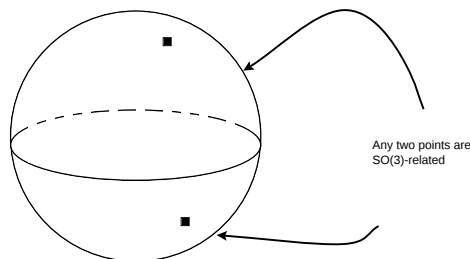


Figure 50: Transitive action of $SO(3, \mathbb{R})$ on the sphere.

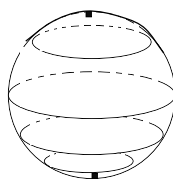


Figure 51: Orbits of $SO(2, \mathbb{R})$ on the two sphere.

7.2 First examples

The concept of a G -action on a set is an extremely important concept, so let us consider a number of examples:

Examples

1. Let $X = \{1, \dots, n\}$, so $S_X = S_n$ as before. The action is effective and transitive, but not free. Indeed, the fixed point of any $j \in X$ is just the permutations that permute everything else, and hence $S_X^j \cong S_{n-1}$. Note that different j have different stabilizer subgroups isomorphic to S_{n-1} , but they are all conjugate.

Now use the stabilizer-orbit theorem to write $|G^x| = |G|/|\mathcal{O}_G(x)|$. Now in the sum

$$\sum_{x \in X} \frac{1}{|\mathcal{O}_G(x)|} \tag{7.14}$$

the contribution of each distinct orbit is exactly 1.

⁵²Hint: Note that $X = G/H$ for some H and apply the Burnside lemma.

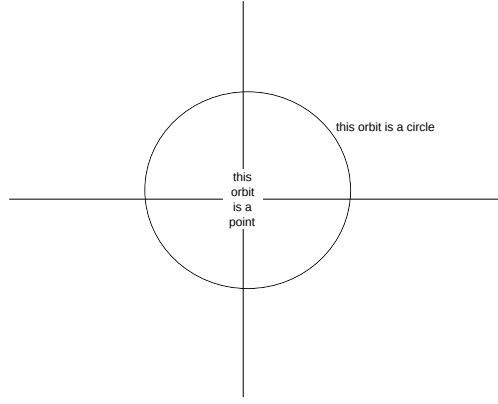


Figure 52: Notice not all orbits have the same dimensionality. There are two qualitatively different kinds of orbits of $SO(2, \mathbb{R})$.

2. *Group actions on the plane.* The group $G = GL(2, \mathbb{R})$ acts on the plane $X = \mathbb{R}^2$ by linear transformation. The action is effective. There are two orbits $O_G(\vec{x})$ depending on whether $\vec{x}$ is zero or not. The action is therefore not transitive, and not free. We can restrict the action to the subgroup $G = SO(2, \mathbb{R})$. The action is:

$$R(\phi) : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \rightarrow \begin{pmatrix} \cos \phi & \sin \phi \\ -\sin \phi & \cos \phi \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad (7.15)$$

The group action is effective. It is not free, and it is not transitive. There are now infinitely many orbits of $SO(2)$, and they are all distinguished by the invariant value of $x^2 + y^2$ on the orbit. From the viewpoint of topology, there are two distinct “kinds” of orbits acting on $\mathbb{R}^2$. One has trivial isotropy group and one has isotropy group $SO(2)$. See Figure 52. These give two strata.

3. Orbits of $O(2)$. The two-dimensions orthogonal group $O(2, \mathbb{R})$ can be written as a semidirect product

$$O(2) = SO(2) \rtimes \mathbb{Z}_2 \quad (7.16)$$

where $\mathbb{Z}_2$ acts on $SO(2)$ by taking $R(\theta) \rightarrow R(-\theta)$. The group has two components which can be written as

$$O(2) = SO(2) \amalg P \cdot SO(2) \quad (7.17)$$

where P is not canonical and can be taken to be reflection in any line through the origin. The orbits of $SO(2)$ and $O(2)$ are the same.

4. Similarly, $SO(3, \mathbb{R})$ acts on $X = \mathbb{R}^3$. It is effective, not transitive, and not fixed-point-free. We can restrict the action to a sphere of any radius S_R^2 . The action is then transitive on the sphere, The isotropy group of any point $x \in S_R^2$ is the subgroup of rotations about the axis through that point. That subgroup is isomorphic to $SO(2, \mathbb{R})$, but as x varies the particular subgroup varies. For example, with usual conventions, if x is on the x^3 -axis then the subgroup is the subgroup of matrices of

the form

$$\begin{pmatrix} \cos \phi & \sin \phi & 0 \\ -\sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (7.18)$$

but if x is on the x^1 -axis the subgroup is the subgroup of matrices of the form

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \phi & \sin \phi \\ 0 & -\sin \phi & \cos \phi \end{pmatrix} \quad (7.19)$$

and so on. There are two strata: Those with G^x congruent to $SO(2, \mathbb{R})$ and those with $G^x = SO(3, \mathbb{R})$.

5. By contrast consider a fixed $SO(2, \mathbb{R})$ subgroup of $SO(3, \mathbb{R})$ defined by rotations around the z-axis. This subgroup also acts on the sphere - but not transitively. The G -orbits are shown in 51.
6. Similarly, $GL(n, \mathbb{R})$ acts on $\mathbb{R}^n$. If we act with a matrix on a column vector we get a left action. If we act on a row vector we get a right action.
7. In general, if G acts on a vector space V over a field κ through κ -linear transformations, that is, for each $g \in G$ the map $v \mapsto \phi(g, v)$ is a linear transformation of V , then V together with the G action constitutes a *representation of G* . Put differently, a G -representation is a vector space V together with a group homomorphism $\rho : G \rightarrow GL(V)$. The subject of representation theory is large and important. See Chapter **** below.
8. If $G = \mathbb{Z}_2$ acts linearly on $\mathbb{R}^{n+1}$ (i.e. $V = \mathbb{R}^{n+1}$ is a representation of $\mathbb{Z}_2$) then we can choose coordinates so that the nontrivial element $\sigma \in G$ acts by

$$\sigma \cdot (x^1, \dots, x^{n+1}) = (x^1, \dots, x^p, -x^{p+1}, \dots, -x^{p+q}) \quad (7.20)$$

where $p + q = n + 1$. Note that this action preserves the equation of the sphere $\sum_i (x^i)^2 - 1 = 0$ and hence descends to a $\mathbb{Z}_2$ -action on the sphere S^n . The case $p = 0, q = n + 1$ is the antipodal map, but there are many other natural actions of $\mathbb{Z}_2$ on S^n .

9. Let the group be $G = \mathbb{C}^*$. Consider a set of integers $(q_1, \dots, q_n) \in \mathbb{Z}^n$. Then for each such set of integers there is a $\mathbb{C}^*$ -action on $\mathbb{CP}^{n-1}$ defined by

$$\mu \cdot [X^1 : \dots : X^n] := [\mu^{q_1} X^1 : \dots : \mu^{q_n} X^n] \quad (7.21)$$

for $\mu \in \mathbb{C}^*$. (Check it is well-defined!)

10. The group $G = SL(2, \mathbb{R})$ acts on the complex upper half plane:

$$\mathcal{H} = \{\tau | \text{Im} \tau > 0\} \quad (7.22)$$

via

$$g \cdot \tau := \frac{a\tau + b}{c\tau + d} \quad (7.23)$$

where

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad (7.24)$$

11. *Actions of $\mathbb{Z}$.* Let us consider $\mathbb{Z}$ to be the free group with one generator g_0 . Then, given any invertible map $f : X \rightarrow X$ we can define a group action of $\mathbb{Z}$ on X by

$$g_0^n \cdot x = \begin{cases} \underbrace{f \circ \dots \circ f}_n(x) & n > 0 \\ x & n = 0 \\ \underbrace{f^{-1} \circ \dots \circ f^{-1}}_{|n|}(x) & n < 0 \end{cases} \quad (7.25)$$

Conversely, any $\mathbb{Z}$ -action must be of this form since we can define $f(x) := g_0 \cdot x$.

12. Let G be any group and consider the group action defined by $\phi(g, x) = x$ for all $g \in G$. This is as ineffective as a group action can be: For every x , the isotropy group is all of G , and for all $g \in G$, $\text{Fix}(g) = X$. In particular, this situation will arise if X consists of a single point. This example is not quite as stupid as might at first appear, once one takes the categorical viewpoint, for $pt//G$ is a very rich category indeed.

Exercise

Consider the action of $\mathbb{Z}_2$ on the sphere defined by (7.20).

- For which values of p, q is the action effective?
- For which values of p, q is the action transitive?
- Compute the fixed point set of the nontrivial element $\sigma \in \mathbb{Z}_2$.
- For which values of p, q is the action free?

Exercise

Consider the action of $G = \mathbb{C}^*$ on $\mathbb{CP}^{n-1}$ defined by (7.21).

- For which values of $(q_1, \dots, q_n)$ is the action effective?
- For which values of $(q_1, \dots, q_n)$ is the action transitive?
- What are the fixed points of the $\mathbb{C}^*$ action?
- What are the stabilizers at the fixed points of the $\mathbb{C}^*$ action?

Exercise

- a.) Show that (7.23) above defines a left-action of $SL(2, \mathbb{R})$ on the complex upper half-plane.⁵³
- b.) Is the action effective?
- c.) Is the action transitive?
- d.) Which group elements have fixed points?
- e.) What is the isotropy group of $\tau = i$?
-

7.3 Action of a topological group on a topological space

If the group G is a topological group it is said to act continuously on a topological space X when $\phi : G \times X \rightarrow X$ is a continuous map. When working with topological groups acting on topological spaces, this is generally assumed. Note that $\phi_g : X \rightarrow X$ has a continuous inverse, namely $\phi_{g^{-1}}$ and therefore ϕ_g is a homeomorphism. Therefore, a topological group action on a topological space can be defined to be a homomorphism from the group G to the group of homeomorphisms on X .⁵⁴ One often wants to work with *proper* actions: This means that the ϕ is a proper continuous map.

Warning: There is some very important, but regrettably very confusing terminology associated with topological group actions on topological spaces. When G is a discrete group there is an important notion of a *properly discontinuous* action. The general definition is that the map $G \times X \rightarrow X \times X$ given by $(g, x) \mapsto (g \cdot x, x)$ is a proper continuous map. This does not mean the function taking $x \mapsto g \cdot x$ is discontinuous! On the contrary, as stated above, it is continuous. To make matters worse, one will find inequivalent definitions of the term “properly discontinuous” in textbooks and on the internet. The subtleties melt away when G is finite or when X is locally compact. We will follow the definitions used by W.P. Thurston, since he was one of the great masters of the subject. Specifically, Definition 3.5.1 of W.P. Thurston, *Three Dimensional Geometry and Topology*, vol 1, Princeton University Press 1997 includes:

Definition: Let G be a discrete group acting continuously on a topological space X . Then

1. The action *has discrete orbits* if every $x \in X$ has a neighborhood U such that the set of group elements $g \in G$ with $g \cdot x \in U$ is finite.
2. The action *is wandering* if every $x \in X$ has a neighborhood U such that the set of group elements $g \in G$ with $g \cdot U \cap U \neq \emptyset$ is finite.
3. If X is locally compact then the action is said to be *properly discontinuous* if for every compact set $K \subset X$ the set of g with $g \cdot K \cap K \neq \emptyset$ is finite.

⁵³Hint: Show that $\text{Im}(g \cdot \tau) = \frac{\text{Im}\tau}{|c\tau + d|^2}$.

⁵⁴Continuous, with a suitable topology on the group of homeomorphisms of X .

If $\mathcal{M}$ is a measure space then a *discrete dynamical system* is a pair of $\mathcal{M}$ together with a measure-preserving group action of $\mathbb{Z}$ on $\mathcal{M}$. According to (7.25) this means we have a map $f : \mathcal{M} \rightarrow \mathcal{M}$ such that $\mu(f(\mathcal{A})) = \mu(\mathcal{A})$ for any measurable set $\mathcal{A}$. For example, in Hamiltonian dynamics one can take $\mathcal{M}$ to be phase space equipped with a symplectic form ω . The natural measure is then the Liouville measure:

$$\mu(\mathcal{A}) := \int_{\mathcal{A}} \frac{\omega^n}{n!} \quad (7.26)$$

In particular, if f is symplectic, i.e. $f^*(\omega) = \omega$, or, in equations:

$$\omega_{\mu\nu}(f(x)) \frac{\partial f^\mu}{\partial x^\lambda} \frac{\partial f^\nu}{\partial x^\rho} = \omega_{\lambda\rho}(x) \quad (7.27)$$

then the corresponding $\mathbb{Z}$ -action is a dynamical system.

One important result is the

Theorem [Poincaré recurrence theorem]. If $f : \mathcal{M} \rightarrow \mathcal{M}$ is a measure-preserving map and has bounded orbits then in any open U set there are points x such that for infinitely many n , $g^n x \in U$.

The proof is based on the idea that if this were not true then the volume of $\cup f^n(U)$ would be infinite, but that cannot be for a volume preserving map with bounded orbits.

There are various ways of expressing how “chaotic” a map is. A dynamical system is said to be *mixing* if for all pairs of (measurable) sets $\mathcal{A}, \mathcal{B} \subset \mathcal{M}$ we have

$$\lim_{n \rightarrow \infty} \mu(\mathcal{A}_n \cap \mathcal{B}) = \frac{\mu(\mathcal{A})\mu(\mathcal{B})}{\mu(\mathcal{M})} \quad (7.28)$$

where $\mathcal{A}_n = f^n(\mathcal{A})$. If $\mu(\mathcal{B}) \neq 0$ this means

$$\lim_{n \rightarrow \infty} \frac{\mu(\mathcal{A}_n \cap \mathcal{B})}{\mu(\mathcal{B})} = \frac{\mu(\mathcal{A})}{\mu(\mathcal{M})} \quad (7.29)$$

so that the “weight” of the set $\mathcal{A}$ is equally distributed over any $\mathcal{B}$. We will give an example of a dynamical system which is mixing below.

Let us examine some special cases of discrete dynamical systems:

1. First, take $X = \mathbb{R}^2$ and let

$$f(x_1, x_2) = (\lambda x_1, \lambda^{-1} x_2) \quad (7.30)$$

where λ is a positive real number greater than 1. For $n > 0$, g_0^n stretches in the x_1 direction and flattens in the x_2 direction. For $n < 0$ the situation is reversed. This action clearly does *not* have discrete orbits, since the origin $(0, 0)$ is a fixed point for the entire group. This action can come up, for example in symplectic geometry: Note that this is a symplectic action with Poisson bracket $\{x_1, x_2\} = 1$, i.e. symplectic form $\omega = dx_1 \wedge dx_2$. It also comes up in some string theory models of cosmological singularities, where we view x_1, x_2 as light-cone coordinates in $1 + 1$ -dimensional Minkowski space.

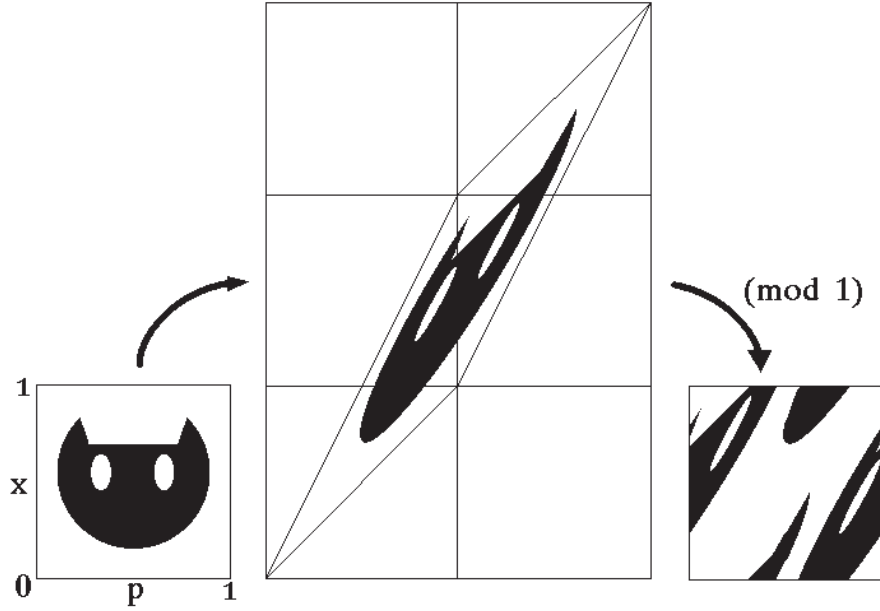


Figure 53: The famous Arnold cat map. The picture ultimately comes from the book V. Arnold and A. Avez, *Ergodic Problems in Classical Mechanics*.

2. Next, consider $X = \mathbb{R}^2 - \{0, 0\}$. Now we can consider (7.30) to be a symplectic action with Poisson bracket for $\{x_1, x_2\} = 1$ just by restriction. However, having excised the point $(0, 0)$ we are now free to define the symplectic structure $\{x_1, x_2\} = x_1 x_2$, i.e.

$$\omega = \frac{dx_1}{x_1} \wedge \frac{dx_2}{x_2} \quad (7.31)$$

and since $x_1 x_2$ is preserved the action (7.30) is still symplectic. Now the action has discrete orbits and is in fact wandering: Consider a neighborhood of some point. We may assume the projection on the x_1 axis will be an open interval (a, b) . Then there are only a finite number of solutions to $a < \lambda^k b < b$ and only a finite number of solutions to $a < \lambda^k a < b$. On the other hand, the action is *not* properly discontinuous. To see this, consider the compact set K which is a closed line segment from $(1, 0)$ to $(0, 1)$. For any $n \neq 0$ there is a solution to

$$\begin{aligned} x + y &= 1 \\ \lambda^n x + \lambda^{-n} y &= 1 \end{aligned} \quad (7.32)$$

with $0 < x < 1$ and $0 < y < 1$. Therefore $g_0^n \cdot K \cap K \neq \emptyset$ for all n .

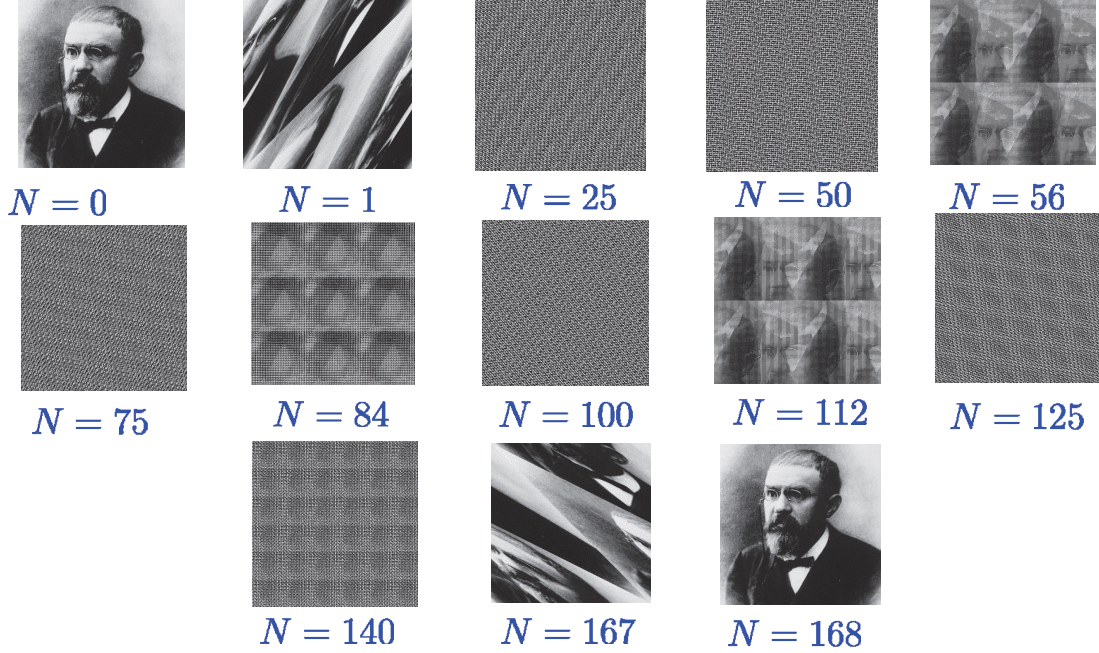


Figure 54: Illustrating Poincaré recurrence for the discrete cat map with $N \times N$ pixels where $N = 294$. Thanks to Andrew Moore for writing the code.

3. *The Arnold cat map:* Consider the torus as $T^2 = [0, 1]^2 / \sim$ with $x \sim x + 1$ and $y \sim y + 1$. Consider the transformation $f : T^2 \rightarrow T^2$ defined by

$$f : \begin{pmatrix} x \\ y \end{pmatrix} \rightarrow \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (7.33)$$

Note again that this is a symplectic transformation. This is known as *Arnold's cat map*, and is famous in dynamical systems theory and in discussions of chaos. Figure 53 shows one iteration of this map. It is shown in Arnold-Avez that the map is mixing.

4. The cat map can also be used to give a dramatic illustration of Poincaré recurrence. A computer screen will have a finite number of pixels. Let us say it has $N \times N$ pixels. If these are black or white there will be $2^{N \times N}$ possible images.⁵⁵ Now, consider the discrete version of the cat map: We take

$$x, y \in \frac{1}{N} \{0, 1, \dots, N - 1\} \quad (7.34)$$

⁵⁵ Actually, there are 256 shades of grey, so actually there are 2^{8N^2} images on a typical computer screen.

and apply the above transformation. Since there are only $2^{N \times N}$ images the transformation must be periodic. See Figure 54 for $N = 294$. It is rather astonishing that the period is only 164. Due to some special number-theoretic aspects of this example (related to Fibonacci numbers) one can give an exact formula for a period (not necessarily the fundamental period), which in this case turns out to confirm $N = 164$. Using this formula it turns out that the period is $\leq 3N$.⁵⁶

5. In general, suppose that a $\mathbb{Z}$ -action on a manifold M is generated by a differentiable map $g_0 \cdot x = f(x)$, where $f : M \rightarrow M$, and suppose that f has a fixed-point $f(x_0) = x_0$. Then, near the fixed point we can write, in local coordinates:

$$f(x_0 + \delta x) = f(x_0) + df(\delta x) = x_0 + df(\delta x) + \mathcal{O}((\delta x)^2) \quad (7.35)$$

where we identify an infinitesimal deviation with a tangent vector. In general $df : T_{x_0}X \rightarrow T_{x_0}X$ is just a real linear transformation and cannot be diagonalized. Suppose, however, that it can be diagonalized. Then there is a basis of $T_{x_0}X$ such that df has the matrix representation $A = \text{Diag}\{\alpha_1, \dots, \alpha_n\}$ for an n -dimensional manifold. If we consider orbits that begin close to x_0 then choose coordinates so that $x_0 = 0 \in \mathbb{R}^n$ and we have $x_{n+1} \cong (1 + A)x_n$. In the directions with $\alpha_j < 0$ the orbits contract to the fixed point. In the directions $\alpha_j > 0$ they expand away from the fixed point (and soon go beyond the linear approximation). If $\alpha_j = 0$ then we need to go to higher order to determine if the fixed point is isolated.

Remark: A very important dynamical system for quantum field theory is known as the “renormalization group.” It has both discrete and continuous forms. One of the discrete forms is known as the block spin method. The dynamical system evolves on the infinite-dimensional space of all possible local couplings of the field theory. Fixed point loci are known as “scale invariant theories,” and they are usually conformal field theories. One then defines local quantum field theory by a scaling of couplings near a fixed point. The key miracle is that, in this infinite dimensional space of couplings all but finitely many directions are attractive (“irrelevant operators”) and only a finite number of directions are repulsive (“relevant operators”) so, after a finite number of choices one has predictive power, at least for renormalizable quantum field theories.

Exercise

Consider an action of a discrete group on a topological space X . Show that properly discontinuous implies wandering implies discrete orbits.

⁵⁶F.J. Dyson and H. Falk, “Period of a Discrete Cat Mapping,” Amer. Math. Monthly, vol. 99 (1992), pp.603-614

Exercise

- a.) Show that the Arnold cat map is a product of a shear by one unit in the x -direction followed by a shear by one unit in the y -direction.
b.) Find a square root of the Arnold cat map.
-

7.4 Left and right group actions of G on itself

Let G be any group. Then G acts on itself as a transformation group, in several ways.

To define the *left action of G on G* we associate to each $a \in G \rightarrow$ the *mapping* $L(a)$, $L(a) : G \rightarrow G$ defined by

$$L(a) : g \mapsto ag \quad (7.36)$$

where on the RHS we use the group multiplication of G . This mapping is 1-1 and onto, so $L(a) \in S_G$.

These transformations satisfy:

$$L(ab) = L(a)L(b) \quad (7.37)$$

and moreover $L(1_G)$ is the identity mapping.

Thus we have defined a left-action of G on itself. It is effective, free, and transitive.

Indeed, we can define a map $\mathcal{L} : a \mapsto L(a)$ which is a homomorphism $\mathcal{L} : G \rightarrow S_G$. Moreover,

$$L(a) = 1 \iff a = 1 \quad (7.38)$$

and hence $\ker \mathcal{L} = \{1_G\}$, so the image of $\mathcal{L}$ is isomorphic to G . Thus, we have proved

Theorem 7.1 (Cayley's Theorem): Any group G is isomorphic to a subgroup of the full permutation group S_G . If $n = |G| < \infty$ then G is isomorphic to a subgroup of S_n .

Warning: For a fixed a , although $L(a)$ is a map $G \rightarrow G$, it is *not* a homomorphism! Note, for example that $L(a)$ takes 1_G to a . Do not confuse this with the fact that the mapping $\mathcal{L} : a \mapsto L(a)$ is a homomorphism.

All of this can be repeated for right-actions: For $a \in G$ define the *right-translation operator* $R(a) : G \rightarrow G$ by $R(a) : g \mapsto g \cdot a$.

Then $\phi_a = R(a)$ defines a right-action of G on itself, and hence $\tilde{\phi}_a = R(a^{-1})$ defines a left-action of G on itself.

It should be fairly obvious that

$$R(a)L(b) = L(b)R(a) \quad (7.39)$$

for all $a, b \in G$. Thus, there is a left $G \times G$ action on G defined by:

$$\phi_{(g_1, g_2)} := L(g_1)R(g_2^{-1}) \quad (7.40)$$

Exercise

Since there is a left-action of $G \times G$ on $X = G$ there is a left-action of the diagonal subgroup $\Delta \subset G \times G$ where $\Delta = \{(g, g) | g \in G\}$ is a subgroup isomorphic to G .

- a.) Show that this action is given by $a \mapsto I(a)$, where $I(a)$ is the conjugation by a . (See, Chapter 1, Section ***)
 - b.) Show that the orbits of Δ are the conjugacy classes of G .
 - c.) What is the stabilizer subgroup of an element $g_0 \in G$?
-

7.5 Induced group actions on function spaces

Let X be a G -set and let Y be any set. There are natural left- and right- actions on the function space $\text{Map}(X, Y)$. Given $\Psi \in \text{Map}(X, Y)$ and $g \in G$ we need to produce a new function $\phi(g, \Psi) \in \text{Map}(X, Y)$. The rules are as follows:

1. If G is a left-action on X then

$$\phi(g, \Psi)(x) := \Psi(g \cdot x) \quad \text{right action on } \text{Map}(X, Y) \quad (7.41)$$

2. If G is a right-action on X then

$$\phi(g, \Psi)(x) := \Psi(g^{-1} \cdot x) \quad \text{left action on } \text{Map}(X, Y) \quad (7.42)$$

3. If G is a left-action on X then

$$\phi(g, \Psi)(x) := \Psi(x \cdot g) \quad \text{left action on } \text{Map}(X, Y) \quad (7.43)$$

4. If G is a right-action on X then

$$\phi(g, \Psi)(x) := \Psi(x \cdot g^{-1}) \quad \text{right action on } \text{Map}(X, Y) \quad (7.44)$$

Example: Consider a spacetime $\mathcal{S}$. With suitable analytic restrictions the space of scalar fields on $\mathcal{S}$ is $\text{Map}(\mathcal{S}, \kappa)$, where $\kappa = \mathbb{R}$ or $\mathbb{C}$ for real or complex scalar fields. If a group G acts on the spacetime, there is automatically an induced action on the space of scalar fields. To be even specific, suppose $X = \mathbb{M}^{1,d-1}$ is d -dimensional Minkowski space time, G is the Poincaré group, and $Y = \mathbb{R}$. Given one scalar field Ψ and a Poincaré transformation $g^{-1} \cdot x = \Lambda x + v$ we have $(g \cdot \Psi)(x) = \Psi(\Lambda x + v)$.

Similarly, suppose that X is any set, but now Y is a G -set. Then again there is a G -action on $\text{Map}(X, Y)$:

$$(g \cdot \Psi)(x) := g \cdot \Psi(x) \quad \text{or} \quad \Psi(x) \cdot g \quad (7.45)$$

according to whether the G action on Y is a left- or a right-action, respectively. These are left- or right-actions, respectively.

We can now combine these two observations and get the general statement: We assume that both X is a G_1 -set and Y is a G_2 -set. We can assume, without loss of generality, that we have left-actions on both X and Y . Then there is a natural $G_1 \times G_2$ -action on $\text{Map}(X, Y)$ defined by:

$$\phi((g_1, g_2), \Psi)(x) := g_2 \cdot (\Psi(g_1^{-1} \cdot x)) \quad (7.46)$$

note that if one writes instead $g_2 \cdot (\Psi(g_1 \cdot x))$ on the RHS then we do not have a well-defined $G_1 \times G_2$ -action (if G_1 and G_2 are both nonabelian). In most applications X and Y both have a G action for a single group and we write

$$\phi(g, \Psi)(x) := g \cdot (\Psi(g^{-1} \cdot x)) \quad (7.47)$$

This is a special case of the general action (7.46), with $G_1 = G_2 = G$ and specialized to the diagonal $\Delta \subset G \times G$.

Example: Again let $X = \mathbb{M}^{1,d-1}$ be a Minkowski space time. Take $G_1 = G_2$ and let $G = \Delta \subset G \times G$ be the diagonal subgroup, and take G to be the Poincaré group. Now let $Y = V$ be a finite-dimensional representation of the Poincaré group. Let us denote the action of $g \in G$ on V by $\rho(g)$. Then a field $\Psi \in \text{Map}(X, Y)$ has an action of the Poincaré group defined by

$$g \cdot \Psi(x) := \rho(g)\Psi(g^{-1}x) \quad (7.48)$$

This is the standard way that fields with nonzero “spin” transform under the Poincaré group in field theory. As a very concrete related example, consider the transformation of electron wavefunctions in nonrelativistic quantum mechanics. The electron wavefunction is governed by a two-component function on $\mathbb{R}^3$:

$$\Psi(\vec{x}) = \begin{pmatrix} \psi_+(\vec{x}) \\ \psi_-(\vec{x}) \end{pmatrix} \quad (7.49)$$

Then, suppose $G = SU(2)$. Recall there is a surjective homomorphism $\pi : G \rightarrow SO(3)$ defined by $\pi(u) = R$ where

$$u\vec{x} \cdot \vec{\sigma} u^{-1} = (R\vec{x}) \cdot \vec{\sigma} \quad (7.50)$$

Then the (double-cover) of the rotation group acts to define the transformed electron wavefunction $u \cdot \Psi$ by

$$(u \cdot \Psi)(\vec{x}) := u \begin{pmatrix} \psi_+(R^{-1}\vec{x}) \\ \psi_-(R^{-1}\vec{x}) \end{pmatrix} \quad (7.51)$$

In particular, $u = -1$ acts trivially on $\vec{x}$ but nontrivially on the wavefunction.

7.5.1 Application: Functions on groups

As a very nice example of the idea of how group actions on a space induce group actions on the functions on that space we touch briefly on the Peter-Weyl theorem and the idea of induced representations.

We first describe the Peter-Weyl theorem:

Let G be a group. Then there is a left action of $G \times G$ on G : $(g_1, g_2) \mapsto L(g_1)R(g_2^{-1})$. Now let $Y = \mathbb{C}$. Then $\text{Map}(G, \mathbb{C})$ is known as the *regular representation* of G because the induced left-action:

$$((g_1, g_2) \cdot \Psi)(h) := \Psi(g_1^{-1}hg_2) \quad (7.52)$$

converts the vector space of functions $\Psi : G \rightarrow \mathbb{C}$ into a representation space for $G \times G$. Suppose, on the other hand that V is a linear representation of G . As mentioned above this means we have a group homomorphism $\rho : G \rightarrow \text{End}(V)$. Then consider the vector space of linear transformations $\text{End}(V)$ of V to itself. This is also a representation of $G \times G$ because if $T \in \text{End}(V)$ then we can define a linear left-action of $G \times G$ on $\text{End}(V)$ by:

$$(g_1, g_2) \cdot T := \rho(g_1) \circ T \circ \rho(g_2)^{-1} \quad (7.53)$$

Now, we have two representations of $G \times G$. How are they related? If V is finite-dimensional we have a map

$$\iota : \text{End}(V) \rightarrow \text{Map}(G, \mathbb{C}) \quad (7.54)$$

The map ι takes a linear transformation $T : V \rightarrow V$ to the complex-valued function $\Psi_T : G \rightarrow \mathbb{C}$ defined by

$$\Psi_T(g) := \text{Tr}_V(T\rho(g^{-1})) \quad (7.55)$$

If we choose a basis w_μ for V then the operators $\rho(g)$ are represented by matrices:

$$\rho(g) \cdot w_\nu = \sum_\mu D(g)_{\mu\nu} w_\mu \quad (7.56)$$

If we take $T = e_{\nu\mu}$ to be the matrix unit in this basis then Ψ_T is the function on G given by the matrix element $D(g^{-1})_{\mu\nu}$. So the Ψ_T 's are linear combinations of matrix elements of the representation matrices of G . The advantage of (7.55) is that it is completely canonical and basis-independent.

Note that $\iota : T \mapsto \Psi_T$ “commutes with the $G \times G$ action.” What this means is that

$$(g_1, g_2) \cdot \Psi_T = \Psi_{(g_1, g_2) \cdot T} \quad (7.57)$$

(The reader should check this carefully.) Such a map is said to be *equivariant*. Put differently, denoting by $\rho_{\text{End}(V)}$ the representation of $G \times G$ on $\text{End}(V)$ and $\rho_{\text{Reg.Rep.}}$ the representation of $G \times G$ on $\text{Map}(G, \mathbb{C})$ we get a commutative diagram:

$$\begin{array}{ccc} \text{End}(V) & \xrightarrow{\iota} & \text{Map}(G, \mathbb{C}) \\ \downarrow \rho_{\text{End}(V)} & & \downarrow \rho_{\text{Reg.Rep.}} \\ \text{End}(V) & \xrightarrow{\iota} & \text{Map}(G, \mathbb{C}) \end{array} \quad (7.58)$$

In particular if we have a collection of finite-dimensional representations $\{V_\lambda\}$ of G then we have

$$\oplus_\lambda \text{End}(V_\lambda) \hookrightarrow \text{Map}(G, \mathbb{C}) \quad (7.59)$$

Thanks to the equivariance, the image of (7.59) is a $G \times G$ -invariant subspace, i.e. a subrepresentation of $\text{Map}(G, \mathbb{C})$. The very beautiful *Peter-Weyl theorem* states that, if G

is a compact group, then, as representations of $G \times G$, (7.59) is an isomorphism if the sum is over the distinct isomorphism classes λ of irreducible representations V_λ of G and we restrict to the subspace of $\text{Map}(G, \mathbb{C})$ of L^2 -normalizable functions with respect to a left-right-invariant measure on G . See Chapter ***** below for a detailed discussion.⁵⁷ For now, we just content ourselves with the statement of the theorem for G a finite group:

Theorem: Let G be a finite group, and define an Hermitian inner product on $L^2(G) = \text{Map}(G, \mathbb{C})$ by

$$(\Psi_1, \Psi_2) := \frac{1}{|G|} \sum_g \Psi_1^*(g) \Psi_2(g) \quad (7.60)$$

Then let $\{V_\lambda\}$ be a set of representatives of the distinct isomorphism classes of irreducible unitary representations for G . For each representation V_λ choose an ON basis $w_\mu^{(\lambda)}$, $\mu = 1, \dots, n_\lambda := \dim_{\mathbb{C}} V_\lambda$. Then the matrix elements $D_{\mu\nu}^\lambda(g)$ defined by

$$\rho(g) w_\nu^{(\lambda)} = \sum_{\mu=1}^{n_\lambda} D_{\mu\nu}^\lambda(g) w_\mu^{(\lambda)} \quad (7.61)$$

form a complete orthogonal set of functions on $L^2(G)$ so that

$$(D_{\mu_1 \nu_1}^{\lambda_1}, D_{\mu_2 \nu_2}^{\lambda_2}) = \frac{1}{n_\lambda} \delta^{\lambda_1, \lambda_2} \delta_{\mu_1, \mu_2} \delta_{\nu_1, \nu_2} \quad (7.62)$$

Idea of proof: The proof is based on linear algebra and Schur's lemma. The normalization constant on the RHS of (7.62) is easily determined by setting $\lambda_1 = \lambda_2$ and $\nu_1 = \nu_2 = \nu$ and summing on ν , and using the hypothesis that these are matrix elements in a *unitary* representation. The relation to (7.59) is obtained by noting that the linear transformations $T = e_{\nu\mu}$, given by matrix units relative to the basis $w_\mu^{(\lambda)}$ form a basis for $\text{End}(V_\lambda)$. ♠

Example 1: Let $G = \mathbb{Z}_2 = \{1, \sigma\}$ with $\sigma^2 = 1$. Then the general complex valued-function on G is specified by two complex numbers $(\psi_+, \psi_-) \in \mathbb{C}^2$:

$$\Psi(1) = \psi_+ \quad \Psi(\sigma) = \psi_- \quad (7.63)$$

This identifies $\text{Map}(G, \mathbb{C}) \cong \mathbb{C}^2$ as a vector space. There are just two irreducible representations $V_\pm \cong \mathbb{C}$ with $\rho_\pm(\sigma) = \pm 1$, because for any representation $\rho(\sigma)$ on a vector space V we can form orthogonal projection operators $P_\pm = \frac{1}{2}(1 \pm \rho(\sigma))$ onto direct sums of the irreps. They are obviously unitary with the standard Euclidean norm on $\mathbb{C}$. The matrix elements give two functions on the group $D^\pm$:

$$D^+(1) = 1 \quad D^+(\sigma) = 1 \quad (7.64)$$

$$D^-(1) = 1 \quad D^-(\sigma) = -1 \quad (7.65)$$

⁵⁷In order for this to be completely correct we need to assume that G is a compact group. Then we introduce a left-right-invariant measure on G and replace the LHS by $L^2(G)$.

(Here and in the next examples when working with 1×1 matrices we drop the $\mu\nu$ subscript!) The reader can check they are orthonormal, and they are complete because any function Ψ can be expressed as:

$$\Psi = \frac{\psi_+ + \psi_-}{2} D^+ + \frac{\psi_+ - \psi_-}{2} D^- \quad (7.66)$$

Example 2: We can generalize the previous example slightly by taking $G = \mathbb{Z}/n\mathbb{Z} = \langle \omega | \omega^n = 1 \rangle$. Let us identify this group with the group of n^{th} roots of unity and choose a generator $\omega = \exp[2\pi i/n]$. Since G is abelian all the representation matrices can be simultaneously diagonalized so all the irreps are one-dimensional. They are:

$V = \mathbb{C}$ and $\rho_m(\omega) = \omega^m$ where m is an integer. Note that $m \sim m + n$ so the set of irreps is again labeled by $\mathbb{Z}/n\mathbb{Z}$ and in fact, under tensor product the set of irreps itself forms a group isomorphic to $\mathbb{Z}/n\mathbb{Z}$.

The matrix elements in the irrep (ρ_m, V) are

$$D^{(m)}(\omega^j) = \omega^{mj} = e^{2\pi i \frac{mj}{n}} \quad (7.67)$$

Now we can check that indeed

$$\frac{1}{|G|} \sum_{g \in G} (D^{(m_1)}(g))^* D^{(m_2)}(g) = \delta_{m_1 - m_2 = 0 \bmod n} \quad (7.68)$$

The decomposition of a function Ψ on the group G is known as the discrete Fourier transform.

Remark: The theorem applies to all compact Lie groups. For example, when $G = U(1) = \{z | |z| = 1\}$ then the invariant measure on the group is just $-i \frac{dz}{z} = \frac{d\theta}{2\pi}$ where $z = e^{i\theta}$:

$$(\Psi_1, \Psi_2) = \int_0^{2\pi} \frac{d\theta}{2\pi} (\Psi_1(\theta))^* \Psi_2(\theta) \quad (7.69)$$

Now, again since G is abelian the irreducible representations are 1-dimensional and the unitary representations are (ρ_n, V_n) where $n \in \mathbb{Z}$, $V_n \cong \mathbb{C}$ and

$$\rho_n(z) := z^n \quad (7.70)$$

Now, the orthonormality of the matrix elements is the standard orthonormality of $e^{in\theta}$ and the Peter-Weyl theorem specializes to Fourier analysis: An L^2 -function $\Psi(\theta)$ on the circle can be expanded in terms of the matrix elements of the irreps:

$$\Psi = \sum_{\text{Irreps } \rho_n} \hat{\Psi}_n D^{(n)} \quad (7.71)$$

When applied to $G = SU(2)$ the matrix elements are known as *Wigner functions* or *monopole harmonics*. They are the matrix elements

$$D_{m_L, m_R}^j(g) := \langle m_L | \rho^j(g) | m_R \rangle \quad (7.72)$$

in the standard ON basis of the unitary spin j representation diagonalizing the diagonal subgroup of $SU(2)$. So

$$j = 0, \frac{1}{2}, 1, \frac{3}{2}, \dots, \quad m_L, m_R \in \{-j, -j+1, \dots, j-1, j\} \quad (7.73)$$

Recall that $SU(2) \cong S^3$ as a manifold. Using the standard volume form, with unit volume we can define $L^2(SU(2))$. The entire theory of spherical harmonics and Legendre polynomials is easily derived from basic group theory.

Now let us turn to induced representations:

Let G be a group and H a subgroup. Suppose that $\rho : H \rightarrow \text{End}(V)$ is a representation of the *subgroup* H . Then, as we have seen $\text{Map}(G, V)$ is canonically a $G \times H$ -space. To keep the notation under control we denote a general function in $\text{Map}(G, V)$ by Ψ . Then the left-action of $G \times H$ defined by declaring that for $(g, h) \in G \times H$ and $\Psi \in \text{Map}(G, V)$ the new function $\phi((g, h), \Psi) \in \text{Map}(G, V)$ is the function $G \rightarrow V$ defined by:

$$\phi((g, h), \Psi)(g_0) := \rho(h) \cdot \Psi(g^{-1}g_0h) \quad (7.74)$$

for all $g_0 \in G$. Now, we can consider the subspace of functions *fixed by the action of* $1 \times H$. That is, we consider the H -equivariant functions which satisfy

$$\boxed{\Psi(gh^{-1}) = \rho(h)\Psi(g)} \quad (7.75)$$

for every $g \in G$ and $h \in H$. Put differently: There are two natural left-actions on $\text{Map}(G, V)$ and we consider the subspace where they are equal. Note that the space of such functions is a linear subspace of $\text{Map}(G, V)$. We will denote it by $\text{Ind}_H^G(V)$. Moreover, it is still a representation of G since if Ψ is equivariant so is $(g, 1) \cdot \Psi$.

The subspace $\text{Ind}_H^G(V) \subset \text{Map}(G, V)$ of H -equivariant functions, i.e. functions satisfying (7.75) is called the *induced representation of G , induced by the representation V of the subgroup H* . This is an important construction with a beautiful underlying geometrical interpretation. In physics it yields:

1. The irreducible unitary representations of space groups in condensed matter physics.
2. The irreducible unitary representations of the Poincaré group in QFT.

Example: Let us take $V = \mathbb{C}$ with the trivial representation of H , i.e. $\rho(h) = 1$. Then the induced representation is the vector space of functions on G which are invariant under right-multiplication by H . This is precisely the vector space of $\mathbb{C}$ -valued functions on the homogeneous space G/H . For example, the invariant Wigner functions D_{m_L, m_R}^j under right-action by the diagonal $U(1)$ subgroup of $SU(2)$ are $D_{m_L, 0}^j(g)$. These descend to functions on $SU(2)/U(1) \cong S^2$ known (for j integral) as the *spherical harmonics*. The case of V a trivial representation generalizes in a beautiful way: When (ρ, V) is nontrivial the induced representation is interpreted not as a space of functions on G/H but rather as a

vector space of sections of a homogeneous vector bundle over G/H determined by the data (ρ, V) . See **** below.

Exercise

Prove (7.57).

Exercise

Let G be the symmetric group on $\{1, 2, 3\}$ and let $H = \{1, (12)\}$. Choose a representation of H with $V \cong \mathbb{C}$ and $\rho(\sigma) = +1$ or $\rho(\sigma) = -1$.

a.) Show that in either case, the induced representation $\text{Ind}_H^G(V)$ is a three-dimensional vector space.

b.) Choose a basis for $\text{Ind}_H^G(V)$ and compute the representation matrices of the elements of S_3 explicitly.

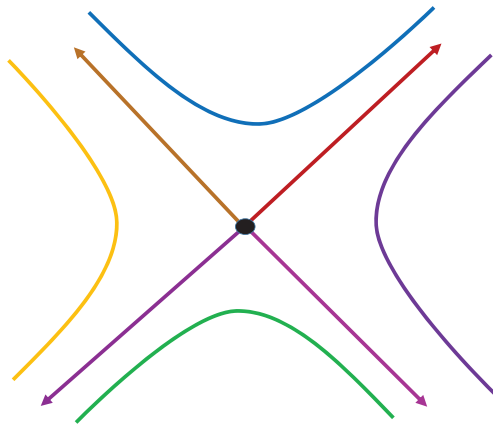


Figure 55: The distinct kinds of orbits of $SO(1, 1, \mathbb{R})$ are shown in different colors. If we enlarge the group to include transformations that reverse the orientation of time and/or space then orbits of the larger group will be made out of these orbits by reflection in the space or time axis.

7.6 An Example of Orbits in physics: Orbits of the Lorentz group and relativistic wave equations

7.6.1 The case of 1 + 1 dimensions

Consider 1+1-dimensional Minkowski space with coordinates $x = (x^0, x^1)$ and metric given by

$$\eta := \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \quad (7.76)$$

i.e. the quadratic form is $(x, x) = -(x^0)^2 + (x^1)^2$. The two-dimensional Lorentz group is defined by

$$O(1, 1) = \{A | A^{tr} \eta A = \eta\} \quad (7.77)$$

This group acts on $\mathbb{M}^{1,1}$ preserving the Minkowski metric.

The connected component of the identity is the group of Lorentz boosts of rapidity θ :

$$t \rightarrow \cosh \theta \, t + \sinh \theta \, x \quad (7.78)$$

$$x \rightarrow \sinh \theta \, t + \cosh \theta \, x \quad (7.79)$$

that is:

$$SO_0(1, 1; \mathbb{R}) \equiv \{B(\theta) = \begin{pmatrix} \cosh \theta & \sinh \theta \\ \sinh \theta & \cosh \theta \end{pmatrix} \mid -\infty < \theta < \infty\} \quad (7.80)$$

In the notation the S indicates we look at the determinant one subgroup and the subscript 0 means we look at the connected component of 1. This is a group since

$$B(\theta_1)B(\theta_2) = B(\theta_1 + \theta_2) \quad (7.81)$$

so $SO_0(1, 1) \cong \mathbb{R}$ as groups. Indeed, note that

$$B(\theta) = \exp \left[\theta \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right] \quad (7.82)$$

It is often useful to define *light cone coordinates*:⁵⁸

$$x^\pm := x^0 \pm x^1 \quad (7.83)$$

and the group action in these coordinates is simply:

$$x^\pm \rightarrow e^{\pm \theta} x^\pm \quad (7.84)$$

so it is obvious that $x^+ x^- = -(x, x)$ is invariant.

⁵⁸Some authors will define these with a $1/2$ or $1/\sqrt{2}$. One should exercise care with this choice of convention.

It follows that the orbits of the Lorentz group are, in general, hyperbolas. They are separated by different values of the Lorentz invariant $x^+x^- = \lambda$, but this is not a complete invariant, since the sign (or vanishing) of x^+ and of x^- is also Lorentz invariant. Define

$$\text{sign}(x) := \begin{cases} +1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} \quad (7.85)$$

Then $(\lambda, \text{sign}(x^+), \text{sign}(x^-))$ is a complete invariant of the orbits. That is, given this triple of data there is a unique orbit with these properties.

It is now easy to see what the different type of orbits are. They are shown in Figure 55: They are:

1. hyperbolas in the forward/backward lightcone and the left/right of the lightcone
2. 4 disjoint lightrays.
3. the origin: $x^+ = x^- = 0$.

It is now interesting to consider the orbits of the full Lorentz group $O(1,1)$ and its relation to the massless wave equations. This group has four components and is, group-theoretically

$$O(1,1) = SO_0(1,1) \rtimes (\mathbb{Z}_2 \times \mathbb{Z}_2) \quad (7.86)$$

We can write, noncanonically,

$$O(1,1) = SO_0(1,1) \amalg P \cdot SO_0(1,1) \amalg T \cdot SO_0(1,1) \amalg PT \cdot SO_0(1,1) \quad (7.87)$$

with

$$P = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad T = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \quad (7.88)$$

The P and T operations map various orbits of $SO_0(1,1)$ into each other: P is a reflection in the time axis and T is a reflection in the space axis. Thus the orbits of the groups $SO(1,1)$, $SO_0(1,1) \amalg PT \cdot SO_0(1,1)$, and $O(1,1)$ all differ slightly from each other.

As an example of a physical manifestation of orbits let us consider the energy-momentum dispersion relation of a particle of mass m with energy-momentum $(E, p) \in \mathbb{R}^{1,1}$.

1. Massive particles: $m^2 > 0$ have (E, p) along an orbit in the upper quadrant:

$$\mathcal{O}^+(m) = \{(m \cosh \theta, m \sinh \theta) | \theta \in \mathbb{R}\} \quad (7.89)$$

2. Massless particles move at the speed of light. In 1+1 dimensions there is an interesting distinction: Rightmoving particles have support on $k_+ = 0$ and $k_- \neq 0$. Leftmoving particles have support on $k_- = 0$.
3. Tachyons have $m^2 < 0$ and have their support on the left or right quadrant. Since k^0 is imaginary for small k^1 some plane waves will have exponential growth. These signal an instability, and some important new physical input is needed.
4. A massless “particle” of zero energy and momentum.

7.6.2 Orbits, Representations, and Differential Equations

Now let us turn to how these orbits are related to some important differential equations in field theory.

As we have seen, since $O(1, 1)$ acts on $\mathbb{M}^{1,1}$, it follows that it acts on the space of fields $\text{Map}(\mathbb{M}^{1,1}, \kappa)$, where $\kappa = \mathbb{R}$ or $\mathbb{C}$ for a real or complex-valued scalar field. For a *scalar* field recall the action of $A \in O(1, 1)$ on the space of solutions is

$$(A \cdot \Psi)(x) := \tilde{\Psi}(x) := \Psi(A^{-1}x) \quad (7.90)$$

Note the A^{-1} in the argument of Ψ in the second equality. This is necessary to get a left-action of the group on the space of fields. If we use Ax then we get a right-action.

Now, quite generally, if V is a representation space for G , and $\mathcal{O} \in \text{End}(V)$ is an invariant linear operator, i.e. an operator which commutes with the action of G ,

$$\rho(g)\mathcal{O} = \mathcal{O}\rho(g) \quad (7.91)$$

then any eigenspace, say $\{v \in V | \mathcal{O}v = \lambda v\}$ will be a sub-representation of G .

Consider the operator

$$\partial^\mu \partial_\mu = -\partial_0^2 + \partial_1^2 \quad (7.92)$$

acting on the space of scalar fields. This is an example of an invariant operator, as one confirms with a simple computation. It can be made manifest by writing

$$\partial^\mu \partial_\mu = -4\partial_+ \partial_- \quad (7.93)$$

where

$$\partial_\pm := \frac{\partial}{\partial x^\pm} = \frac{1}{2}(\partial_0 \mp \partial_1) \quad (7.94)$$

The *Klein-Gordon equation* for a complex or real scalar field $\Psi(x^0, x^1)$ is

$$(\partial^\mu \partial_\mu + m^2) \Psi = 0 \quad (7.95)$$

The space of fields satisfying the KG equation is a representation space of $O(1, 1)$, by our general remark above.

Now we relate the orbits of $O(1, 1)$ to the representations furnished by solutions of the KG equation:

For field configurations which are Fourier transformable we can write

$$\Psi(x^0, x^1) = \int dk_0 dk_1 \hat{\psi}(k) e^{ik_0 x^0 + ik_1 x^1} \quad (7.96)$$

If the field Ψ is on-shell then $\hat{\psi}(k)$ must have support on set

$$\{k : (k_0)^2 - (k_1)^2 = m^2\} \quad (7.97)$$

that is:

$$(k^2 + m^2)\hat{\psi}(k) = 0 \quad (7.98)$$

Thus, the support in Fourier space is an orbit. In physics the orbit with $k_0 > 0$ is called the *mass-shell*.

For example, suppose that $m^2 > 0$. Then on the orbit with $k^0 > 0$ we have $k = m(\cosh \theta, \sinh \theta)$ and we can write the general complex-valued solution as

$$\hat{\psi}(k) = \delta(k^2 + m^2)\bar{a}, \quad (7.99)$$

where the amplitude $\bar{a}$ should be regarded as a complex-valued function on the orbit. Then a complex-valued solution on the KG equation is generated by such a function on the orbit by:

$$\begin{aligned} \Psi(x) &= \int d^2k \delta(k^2 + m^2) \bar{a} e^{ik \cdot x} \\ &= \int_{\mathbb{R}} \frac{dk_1}{2\sqrt{k_1^2 + m^2}} \bar{a}(k_1) e^{ik \cdot x} \\ &= \frac{1}{2} \int_{-\infty}^{\infty} d\theta \bar{a}(\theta) e^{im(x^0 \cosh \theta + x^1 \sinh \theta)} \end{aligned} \quad (7.100)$$

where we chose a convention designation of the argument of the function $\bar{a}$ on the orbit.

Thus, we can identify a space of solutions to the KG equation with a space of functions on an orbit.

Now recall that the space of functions on an orbit is a special case of an induced representation. As we will soon see, induced representations are the right concept for generalization to other representations of the Lorentz group.

Remark 1: *The components of the Lorentz group:* If we use a particular orbit to generate representations we only represent the subgroup of the Lorentz group which preserves that orbit. Thus, for the orbit with $m > 0$ and $k_0 > 0$ we can include the parity-reversing component, but not the time-reversing component. Of course, there is a similar representation of $SO_0(1, 1) \amalg P \cdot SO_0(1, 1)$ given by the hyperbola with $k_0 < 0$. If we wish to represent the full $O(1, 1)$ then we must include both orbits. Note that, if we wish to use *real* solutions of the KG equation we must include both hyperbolae and there is a reality condition relating the Fourier modes: $(\bar{a}(k))^* = a(-k)$

Remark 2: The representation can be made into a unitary representation. See below.

Exercise

Show that an invariant linear operator $\mathcal{O} \in \text{End}(V)$ on a representation space V of G is a fixed point of the $G \times G$ action on $\text{End}(V)$.

7.6.3 The massless case in 1 + 1 dimensions

Having phrased things this way it is natural to ask what relativistic wave equations correspond to amplitudes supported on the light-like orbits.

In the massless case $m^2 = 0$ the general solution of the KG equation is easily written as

$$\Psi = \psi_L(x^+) + \psi_R(x^-) \quad (7.101)$$

Solutions of the form $f_R = 0$ are called *left-moving waves* because, as time evolves forward, the profile of the function f_L moves to the left on the x^1 -axis. Similarly, solutions of the form $f_L = 0$ are called *right-moving waves*.

Remark: In the massless case there is a “more primitive” relativistic wave equation which is first order, and whose solutions are always solutions of the massless KG equation. Namely, we can consider the separate equations

$$\partial_+ \Psi = 0 \quad (7.102)$$

$$\partial_- \Psi = 0 \quad (7.103)$$

Note that these equations are themselves Lorentz invariant, even though the operators $\partial_\pm$ are not invariant differential operators. Solutions to (7.102) are right-moving scalar fields and solutions to (7.103) are left-moving scalar fields. Such on-shell scalar fields are also known as *chiral scalar fields*. Equations (7.102) and (7.103) are notable in part because they play an important role in string theory and conformal field theory. It is also interesting to note that it is quite subtle to write an action principle that leads to such equations of motion. They are also called the anti-self-dual and self-dual equations of motion, respectively because, if we choose the orientation $dx \wedge dt$ then the Hodge star operation is $*dt = dx$ and $*dx = dt$ and hence $*(dx^\pm) = \pm dx^\pm$. Therefore if $\partial_\pm \Psi = 0$ its “fieldstrength” $F = d\Psi$ satisfies $*F = \mp F$.

The action of $P, T \in O(1, 1)$ on a real scalar field is given by:

$$\begin{aligned} (P \cdot \Psi)(x, t) &:= \Psi(-x, t) \\ (T \cdot \Psi)(x, t) &:= \Psi^*(x, -t) \end{aligned} \quad (7.104)$$

The KG equation is separately P and T invariant, but the (anti)-self-dual equations are not. Nevertheless, the latter equations are PT invariant. This is a special case of the famous CPT theorem:

Roughly speaking, mathematical consistency implies that if a physical theory is invariant under group transformations in the neighborhood of the identity then it must be invariant under the transformations in the full connected component of the identity. But this does not mean the theory is invariant under disconnected components such as the components containing P and T . As a matter of fact, Nature chooses exactly that option in the standard model of the electroweak and strong interactions. However, if we also assume that there is a certain relation to “analytically continued” equations in Euclidean signature then, since PT is in the connected component of the identity of $O(2)$, such theories must in fact be PT invariant.

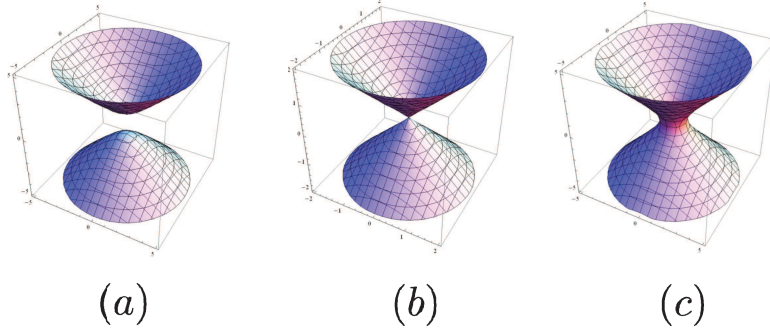


Figure 56: Illustrating orbits of the connected component of the identity in $O(1, 3)$. In (a) the top and bottom hyperboloids are separate orbits, and if we include time-reversing transformations the orbits are unions of the two hyperboloids. In (b) there are three orbits shown with $x^0 > 0$ $x^0 < 0$ (the future and past, or forward and backward light cones), and the orbit consisting of the single point. In (c), once x^2 has been specified, there is just one orbit, for $d > 2$.

7.6.4 The case of d dimensions, $d > 2$

Now consider Minkowski space $\mathbb{M}^{1,d-1}$ with $d > 2$. The nature of the orbits is slightly different.

1. For $\lambda^2 > 0$ we can define

$$\mathcal{O}^+(\lambda) = \{x | (x^0)^2 - (\vec{x})^2 = \lambda^2 \quad \& \quad \text{sign}(x^0) = \text{sign}(\lambda)\} \quad (7.105)$$

By the stabilizer-orbit theorem we can identify this with

$$SO_0(1, d-1)/SO(d-1) \quad (7.106)$$

by considering the isotropy group at $(x^0 = \lambda, \vec{x} = 0)$. See Figure 56(a).

2. For $\mu^2 > 0$ we can define

$$\mathcal{O}^-(\lambda^2) = \{x | (x^0)^2 - (\vec{x})^2 = -\mu^2\} \quad (7.107)$$

By the stabilizer-orbit theorem we can identify this with

$$SO_0(1, d-1)/SO_0(1, d-2) \quad (7.108)$$

by considering the isotropy group at $x = (x^0 = 0, x^1 = 0, \dots, x^{d-2} = 0, x^{d-1} = \mu)$. The sign of μ does not distinguish different orbits for $d > 2$ because the sphere S^{d-2} is connected. See Figure 56(c).

3.

$$\mathcal{O}^\pm = \{x|x^2 = 0 \quad \& \quad \text{sign}(x^0) = \pm 1\} \quad (7.109)$$

Vectors in this orbit are of the form $(x^0, |x^0|\hat{n})$ where $\hat{n} \in S^{d-2} \subset \mathbb{R}^{d-1}$ and the sign of x^0 is invariant under the action of the identity component of $O(1, 3)$. (Show this!). Note that, for $d = 2$ the sphere S^0 has two disconnected components, leading to left- and right-movers. But for $d > 2$ there is only one component. We can think of $\hat{n} \in S^{d-2}$ as parametrizing the directions of light-rays. That is, the point where the light ray hits the celestial sphere. In one spatial dimension, a light ray either moves left or right, and this is a Lorentz-invariant concept. In $d - 1 > 1$ spatial dimensions, we can rotate any direction of light ray into any other. See Figure 56(b). One can show that these orbits too are homogeneous spaces:⁵⁹

$$\mathcal{O}^\pm \cong SO_0(1, d - 1)/\mathcal{I} \quad (7.110)$$

4. The final orbit is of course $\{x = 0\}$.

As in $1 + 1$ dimensions we can identify a representation space of the Lorentz group associated with the space of solutions to the KG equation with functions on various orbits. The formulae are essentially the same. For example for the orbit $\mathcal{O}^+(m)$ a function $\bar{a} : \mathcal{O}^+(m) \rightarrow \mathbb{C}$ determines a solution:

$$\begin{aligned} \Psi(x) &= \int d^d k \delta(k^2 + m^2) \bar{a} e^{ik \cdot x} \\ &= \int_{\mathbb{R}^{d-1}} \frac{d^{d-1} \vec{k}}{2\sqrt{\vec{k}^2 + m^2}} \bar{a}(\vec{k}) e^{ik \cdot x} \end{aligned} \quad (7.111)$$

However, for scalar fields, there is no analog of the left- and right-chiral boson. There *are* analogs involving interesting first order equations such as the Dirac equation and the (anti-) self-dual equations for fields with spin.

Quite generally, we can define an inner product on the space of complex-valued solutions of the KG equation such that the action of the Lorentz group is unitary. Observe that, given any two complex-valued solutions Ψ_1, Ψ_2 the current

$$j_\mu := -i(\Psi_1^* \partial_\mu \Psi_2 - (\partial_\mu \Psi_1)^* \Psi_2) \quad (7.112)$$

Note that, if Ψ_1 and Ψ_2 both satisfy the KG equation then

$$\partial^\mu j_\mu = 0 \quad (7.113)$$

⁵⁹The isotropy group of a light ray is $\mathcal{I} \cong ISO(d - 2)$, where $ISO(d - 2)$ is the Euclidean group on $\mathbb{R}^{d-2}$. The easiest way to show this is to use the Lie algebra of $so(1, d - 1)$ and work with light-cone coordinates. Choosing a direction of the light ray along the x^{d-1} axis and introducing light-cone coordinates $x^\pm := x^0 \pm x^{d-1}$, and transverse coordinates x^i , $i = 1, \dots, d - 2$ if the lightray satisfies $x^- = 0$ then we have unbroken generators M^{+i} and M^{ij} .

is conserved. Therefore, if we choose a spatial slice with normal vector n^μ and induced volume form vol the inner product

$$(\Psi_1, \Psi_2) := \int_{\Sigma} n^\mu j_\mu \text{vol} \quad (7.114)$$

is independent of the choice. So, fixing a Lorentz frame and taking Σ to be the slice at a fixed time we have

$$(\Psi_1, \Psi_2) := -i \int_{\mathbb{R}^{d-1}} (\Psi_1^* \partial_0 \Psi_2 - (\partial_0 \Psi_1)^* \Psi_2) d^{d-1} \vec{x} \quad (7.115)$$

This is clearly not positive definite on the space of all solutions but does become positive definite when restricted to the space of complex solutions associated with a single orbit. Indeed, substituting the expansion in momentum space (7.111) we get

$$(\Psi_1, \Psi_2) = \int_{\mathbb{R}^{d-1}} \frac{d^{d-1} \vec{k}}{2\sqrt{\vec{k}^2 + m^2}} (\bar{a}_1(\vec{k}))^* (\bar{a}_2(\vec{k})) \quad (7.116)$$

Having phrased things this way, it is clear that there is an interesting generalization: We can choose other representations of $H = SO(d-1)$ and consider the induced representations of the Lorentz group. This indeed leads to the unitary representations, corresponding to particles with nontrivial spin.

7.7 Spaces of orbits

We now shift our focus from studying orbits to studying *the space of orbits*.

In general, if G has a right-action on a set X then the set of distinct G -orbits is denoted X/G . If G has a left-action then the set of distinct G orbits is denoted $G \backslash X$.

The study of such spaces of orbits is rather vast. Some of the ways spaces of orbits enter into physics are the following:

1. Spaces of orbits such as homogeneous spaces given beautifully symmetric manifolds (or orbifolds). Thus they form a rich source of geometrical constructions and can be used in constructing spacetimes or discussing moduli spaces of solutions to equations.
2. A natural source of orbits is *Hamiltonian dynamics*. The time evolution of a dynamical system naturally defines a system of $\mathbb{R}$ or $\mathbb{Z}$ -orbits on phase space.
3. If G is a global symmetry group in a field theory, then the vacua of the theory are a union of orbits of G , and hence a union of homogeneous spaces G/H . In general H will be different in different connected components. It is referred to as the *unbroken symmetry* in the theory of spontaneous symmetry breaking.
4. Often, it is convenient to introduce redundant variables in a physical problem to make some other property of the physics (such as locality) manifest. In this case G is a *gauge symmetry* and it acts on the set of redundant variables X while the space of orbits X/G are the physically inequivalent variables. The canonical example

is obtained by identifying X with the set of all gauge potentials and G with the group of gauge transformations. For example, on $\mathbb{M}^{1,3}$ the gauge potentials A_μ are redundant variables. The group $\mathcal{G} = \text{Map}(\mathbb{M}^{1,3}, \mathbb{R})$ acts by $A_\mu \rightarrow A_\mu + \partial_\mu \chi$. The orbit space is parametrized by the fieldstrength $F_{\mu\nu}$.

We cannot cover this topic in proper detail here, but just indicate some examples and definitions to give a taste of the subject.

Remarks

1. As an example, let us consider the case where $X = G$ is itself a group. Let $H \subset G$ be a subgroup and consider the right-action of H on the set X . The set of orbits is G/H . This is in accord with our notation for the set of left-cosets of a subgroup H in a larger group G . Note that the set of orbits still admits a left-action by G . If $K \subset G$ is another subgroup then the set of orbits of the left-action of K on G/H is known as a *double coset* and denoted $K \backslash G/H$.
2. Warning! The notation X/G is somewhat ambiguous, as a space X can admit more than one group action. For example, if $X = G$ and we use right-translation $R(g)$ then X/G is a single point. On the other hand, if $X = G$ and the G action is by conjugation:

$$\phi_g(h) := ghg^{-1} \tag{7.117}$$

then the space X/G is the set of conjugacy classes in G and always has more than one point, if G is nontrivial.

3. When G is a topological group acting on a topological space X we can make X/G into a topological space: We use the quotient topology under the equivalence relation of being G -related. Put differently, the topology on X/G is defined by requiring that $p : X \rightarrow X/G$ be a continuous map. That is, $\mathcal{U} \subset X/G$ is open if $p^{-1}(\mathcal{U})$ is open. Even when X and G are relatively “simple” the quotient can be quite subtle and complicated. For example, even if X and G are Hausdorff the quotient might not be Hausdorff. So, it is useful to have some criteria for when quotients are “well-behaved.”
4. If G is a discrete group we defined the notion of a “properly discontinuous action” above. In this case we have a nice

Theorem: Let G be a discrete group acting freely and properly discontinuously on a Hausdorff manifold X . Then the quotient X/G is a Hausdorff manifold.

To prove this, just note that around each point $x \in X$ there is an open neighborhood U homeomorphic to $\mathbb{R}^n$ such that $g \cdot U \cap U = \emptyset$ for all $g \neq 1$. This U will serve as a local neighborhood. To check that it is Hausdorff show that if x_1, x_2 are on distinct orbits then we can then there are compact neighborhoods K_1 and K_2 of x_1, x_2 which are disjoint. For a more complete proof see Thurston, Proposition 3.5.7.

5. When G is a Lie group one useful criterion is the

Theorem: [Quotient manifold theorem]: Let G be a Lie group acting smoothly, freely, and properly, on a smooth manifold X . Then X/G is a manifold of dimension $\dim X - \dim G$ and the projection map $p : X \rightarrow X/G$ is a submersion.

For a proof see, J. Lee, *Introduction to smooth manifolds*, pp. 218-223.

7.7.1 Simple examples

Example 1 . $G = \mathbb{Z}$ acts properly discontinuously on $\mathbb{R}$ via $n \cdot x = x + n$. The orbits are in 1-1 correspondence with $[0, 1]/\sim$ where $\sim$ identifies $0 \sim 1$. Note that therefore the set of orbits is in one-one correspondence with $\mathbb{R}/\mathbb{Z}$, and can be identified with points on the circle S^1 .

Example 2. This generalizes: Let Λ be a two-dimensional lattice. It acts properly discontinuously on $\mathbb{R}^2$ by translation. Then

$$\mathbb{R}^2/\Lambda \tag{7.118}$$

is geometrically realized as a doughnut, or torus. Of course, nothing is special about two dimensions in this example, if Λ is an n -dimensional lattice in $\mathbb{R}^n$ then it acts on $\mathbb{R}^n$ by translations and the set of orbits is an n -dimensional torus.

Example 3: If we take $X = S^2$ and $G = SO(2)$ acting by rotations around some axis then the space of orbits is the closed interval.

Example 4: If $X = \mathbb{R}^{n+1} - \{0\}$ and $G = \mathbb{R}^*$ acting by scalar multiplication then $X/G \cong \mathbb{RP}^n$. Similarly if $X = \mathbb{C}^{n+1} - \{0\}$ and $G = \mathbb{C}^*$ acting by scalar multiplication then $X/G \cong \mathbb{CP}^n$. In these cases the quotient manifold theorem is not so useful because the groups involved are not compact so it is not easy to check the map is proper. However, if we think of $\mathbb{CP}^n$ as the moduli space of lines through the origin then we can put the standard Hermitian structure on $\mathbb{C}^{n+1}$ and observe that for every line there exists a basis vector v of norm one: $v^* \cdot v = 1$. But

$$\{v \in \mathbb{C}^{n+1} | v^* \cdot v = 1\} \cong S^{2n+1} \tag{7.119}$$

This ON vector is not unique. We can define an action of $U(1)$ on S^{2n+1} by $v \rightarrow e^{i\theta}v$. Then $S^{2n+1}/U(1)$ is a manifold by the quotient manifold theorem, and is another presentation of $\mathbb{CP}^n$. Now that we see that a given space of orbits can be realized as a quotient space in different ways we should note a third presentation: $GL(n+1, \mathbb{C})$ acts on $\mathbb{CP}^n$ by linear action on the homogeneous coordinates. The action is clearly transitive. Therefore, by the stabilizer-orbit theorem we need only find the stabilizer subgroup of a convenient point, such as $[1 : 0 : \dots : 0]$. The stabilizer subgroup of this point is the subgroup B of $GL(n+1, \mathbb{C})$ consisting of matrices of the form:

$$\begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1,n+1} \\ 0_{n \times 1} & \tilde{g} & & \end{pmatrix} \tag{7.120}$$

where $\tilde{g} \in GL(n, \mathbb{C})$ and $g_{11} \in \mathbb{C}^*$. Then $\mathbb{CP}^n$ can be identified with the complex homogeneous space $GL(n+1, \mathbb{C})/B$.

Example 5: If we take $X = \mathbb{R}^2$ with $G = SO_0(1, 1)$ acting via Lorentz transformations then we saw above that the space of orbits X/G is a union of a 5 special points together with four copies of $(0, \infty)$. In this case X/G is not a Hausdorff space. See Section **** below.

7.7.2 Fundamental domains

One way to try to get a picture of the space of orbits of a group action on a space X is to form a *fundamental domain*. In order to define this notion recall that given a G action on X we can define an equivalence relation, saying $x \sim y$ (“ x is G -related to y ” if there is a $g \in G$ with $y = g \cdot x$). In these terms a fundamental domain is a subset $\mathcal{F} \subset X$ of elements providing a complete set of representatives of the distinct equivalence classes for this relation. Then:

$$X = \coprod_{g \in G} g \cdot \mathcal{F}, \quad (7.121)$$

so the points in $\mathcal{F}$ are in one-one correspondence with the points of X/G . Put differently:

Definition: A *fundamental domain* for a group action of G on X is a subset $\mathcal{F} \subset X$ such that:

1. No two distinct points in $\mathcal{F}$ are G -related.
2. Every point of X is G -related to some point in $\mathcal{F}$.

Example 1 . *Finite covers of the circle.* Consider:

$$X = U(1) \cong \{z : |z| = 1\} \cong S^1 \quad (7.122)$$

and let $G \cong \mathbb{Z}/N\mathbb{Z}$ be the subgroup of N^{th} roots of 1. This acts on X by multiplication. A fundamental domain is the set of $z = e^{i\theta}$ with θ varying over a half-open interval of length $2\pi/N$. For example we could take $0 \leq \theta < 2\pi/N$. Then X/G is another copy of S^1 . The map $p : X \rightarrow X/G$ can be identified with the map $p(z) = z^N$.

Example 2 . If we consider $X/G = \mathbb{R}/\mathbb{Z} \cong S^1$ then a fundamental domain would be any set of points $\theta \leq x < \theta + 1$, for any θ . As we see from the previous two examples, fundamental domains are not unique.

Example 3 . If we consider $\mathbb{R}^2/\mathbb{Z}^2$ where $\mathbb{Z}^2$ acts by translations by independent vectors e_1, e_2 then one choice of fundamental domain would be the set of points $\vec{x}_0 + t_1 e_1 + t_2 e_2$ where we take the union

$$\{(t_1, t_2) : 0 < t_1, t_2 < 1\} \amalg \{(0, t_2) : 0 < t_2 < 1\} \amalg \{(t_1, 0) : 0 < t_1 < 1\} \amalg \{(0, 0)\} \quad (7.123)$$

Note that $\vec{x}_0$ is arbitrary, and any choice gives a fundamental domain.

Example 4 . In general *unit cell* $\bar{\mathcal{F}} \subset \mathbb{R}^n$ of an embedded lattice $\Lambda \subset \mathbb{R}^n$ is the closure of a fundamental domain for $\mathbb{R}^n/\Lambda$. One natural choice is given by *choosing* a basis e_i for the lattice together with a vector $\vec{x}_0 \in \mathbb{R}^n$ and defining

$$\bar{\mathcal{F}} = \vec{x}_0 + \left\{ \sum_i t_i e_i \mid 0 \leq t_i \leq 1 \right\}. \quad (7.124)$$

Since $\mathbb{R}^n$ has a natural Euclidean metric the lattice Λ inherits an inner product. In the basis e_i it is given by the symmetric form $G_{ij} := e_i \cdot e_j$. Note that the volume of the unit cell is

$$\text{vol}(\bar{\mathcal{F}}) = \sqrt{\det G_{ij}} \quad (7.125)$$

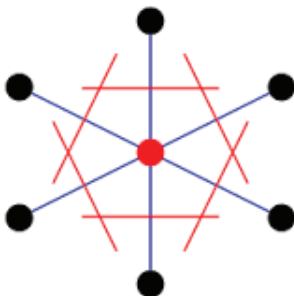


Figure 57: Constructing a Wigner-Seitz (or Voronoi) cell for the triangular lattice. The cells are regular hexagons. Figure from Wikipedia.

Example 5 . Given an embedded lattice $\Lambda \subset \mathbb{R}^n$ we can use the metric to produce a canonical (i.e. basis-independent) set of fundamental domains, known as *Voronoi cells* in mathematics and as *Wigner-Seitz cells* in physics. Choose any lattice point $v \in \Lambda$ and take $\bar{\mathcal{F}}$ to be the set of all points in $\mathbb{R}^n$ which are closer to v than to any other point. (If the points are equidistant to another lattice point we include them in the closure $\bar{\mathcal{F}}$. Thus, for the regular triangular lattice a Wigner-Seitz cell would be a regular hexagon centered on a lattice point. See Figure 57. In reciprocal space, the Wigner-Seitz cell for the reciprocal lattice is known in solid state physics as the *Brillouin zone*. Note that there is a clear algorithm for constructing $\bar{\mathcal{F}}$: Starting with v we look at all other points $v' \in \Lambda$. We consider the hyperplane perpendicular to the line between v and v' and take the intersection of all the half-planes containing v . It is also worth remarking that the

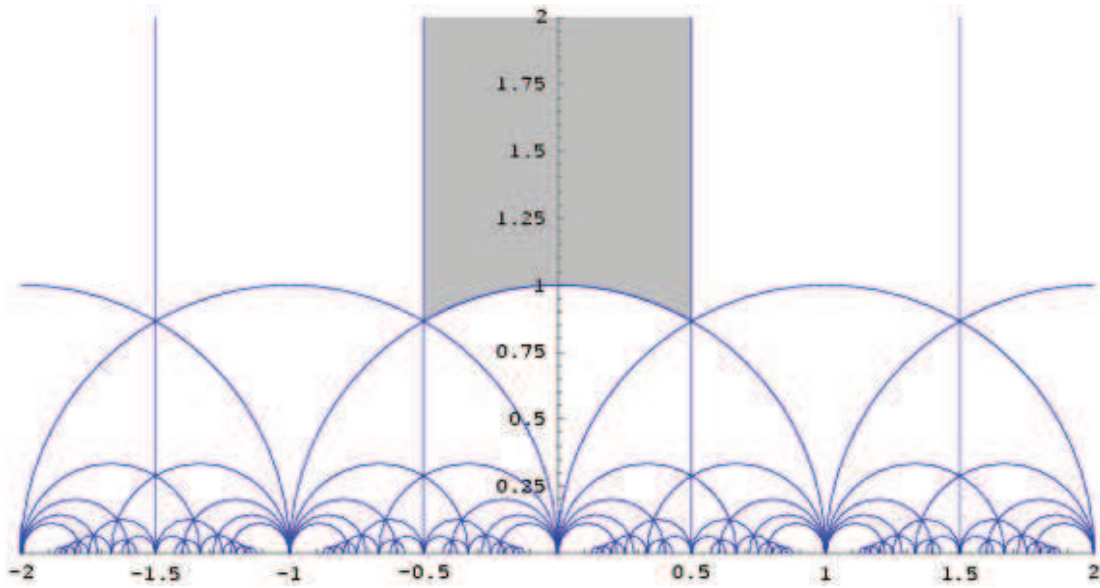


Figure 58: The keyhole region, a standard choice of fundamental domain for the action of $PSL(2, \mathbb{Z})$ on the complex upper half-plane. Figure from Wikipedia article on "Modular Group".

concept of Voronoi cell does not require a lattice and applies to any collection of points, indeed, any collection of subsets of $\mathbb{R}^n$.

Example 6 . The modular group. The modular group is $PSL(2, \mathbb{Z}) := SL(2, \mathbb{Z})/\{\pm 1\}$, where $SL(2, \mathbb{Z})$ is the subgroup of $SL(2, \mathbb{R})$ of matrices all of whose matrix elements are integers. Recall that this group acts effectively on the complex upper half-plane $\mathcal{H}$ via

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \tau := \frac{a\tau + b}{c\tau + d} \quad (7.126)$$

We will find a fundamental domain for this group action, and in the process prove that $SL(2, \mathbb{Z})$ is generated by the group elements S and T defined by:

$$S := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad (7.127)$$

$$T := \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad (7.128)$$

Denote their images in $PSL(2, \mathbb{Z})$ by $\bar{S}, \bar{T}$.⁶⁰

Let:

$$\tilde{\mathcal{F}} := \{\tau \in \mathcal{H} \mid |\tau| \geq 1 \quad \& \quad |\operatorname{Re}(\tau)| \leq \frac{1}{2}\} \quad (7.129)$$

This is almost, but not quite the canonical fundamental domain for the modular group. It is the famous *keyhole region* shown in Figure 58. Let $\bar{G}$ be the subgroup generated by $\bar{S}$ and $\bar{T}$. We claim that $\cup_{g \in \bar{G}} g \cdot \tilde{\mathcal{F}}$ is the entire half-plane. To prove this recall that, for any $g \in SL(2, \mathbb{Z})$,

$$\operatorname{Im}(g \cdot \tau) = \frac{\operatorname{Im} \tau}{|c\tau + d|^2} \quad (7.130)$$

Now, for any fixed $\tau \in \mathcal{H}$ the function $|c\tau + d|$ is bounded below on $SL(2, \mathbb{Z})$, and hence on G . Indeed, decomposing $\tau = x + iy$ into its real and imaginary parts

$$|c\tau + d|^2 = (cx + d)^2 + c^2y^2 \geq \begin{cases} y^2 & c \neq 0 \\ d^2 \geq 1 & c = 0 \end{cases} \quad (7.131)$$

Therefore, for any fixed τ there will exist a group element $g \in \bar{G}$ such that $\operatorname{Im}(g \cdot \tau)$ takes a maximal value as a function of g . Note that multiplying g on the left by a power of $\bar{T}$ or $\bar{T}^{-1}$ does not change this property, so there is not a unique g which maximizes $\operatorname{Im}(g \cdot \tau)$. We can fix the ambiguity by requiring $|\operatorname{Re}(g \cdot \tau)| \leq \frac{1}{2}$. Choose such a group element g . We claim that for this transformation, $\tau' = g \cdot \tau \in \tilde{\mathcal{F}}$. We need only check that $|\tau'| \geq 1$. If not, then $|\tau'| < 1$ but then $\operatorname{Im}(\bar{S} \cdot \tau') = \operatorname{Im}(\tau')/|\tau'|^2 > \operatorname{Im}(\tau')$, contradicting the definition of g . In conclusion, every element of the upper half-plane can be brought to $\tilde{\mathcal{F}}$ by a suitable element of $\bar{G}$.

Now we need two Lemmas:

Lemma 1: If $g \in SL(2, \mathbb{Z})$ and τ have the property that both $\tau \in \tilde{\mathcal{F}}$ and $g \cdot \tau \in \tilde{\mathcal{F}}$ then

1. $|\operatorname{Re}(\tau)| = \frac{1}{2}$ and $g \cdot \tau = \tau \pm 1$, or
2. $|\tau| = 1$

To prove this note that, WLOG, we may assume that $\operatorname{Im}(g \cdot \tau) \geq \operatorname{Im} \tau$. (If not replace $g \rightarrow g^{-1}$ and $\tau \rightarrow g^{-1}\tau$.) But this equation implies $1 \geq |c\tau + d|$ which in turn implies:

$$\begin{aligned} 1 &\geq |c\tau + d|^2 \\ &= (cx + d)^2 + c^2y^2 \\ &= c^2|\tau|^2 + 2cdx + d^2 \\ &\geq c^2|\tau|^2 - |cd| + d^2 \\ &= c^2(|\tau|^2 - \frac{1}{4}) + (|d| - \frac{1}{2}|c|)^2 \\ &\geq \frac{3}{4}c^2 + (|d| - \frac{1}{2}|c|)^2 \end{aligned} \quad (7.132)$$

From (7.132) we conclude:

⁶⁰We are here following a very nice argument by J.-P. Serre, *A Course in Arithmetic*, Springer GTM 7, pp. 78-79.

1. $(c = 0, d = \pm 1)$ or $(c = \pm 1, d = 0, \pm 1)$.
2. The inequalities are saturated iff $2cdx = -|cd|$ and $|\tau| = 1$.

If $c = 0$ and $d = \pm 1$ then $g \cdot \tau = \tau \pm 1$. In this case it is clear that $|\operatorname{Re}(\tau)| = \frac{1}{2}$. If $c = \pm 1$, and $d = 0, \pm 1$ then the inequality is saturated, and hence $|\tau| = 1$.

Lemma 2: If $\tau \in \tilde{\mathcal{F}}$ and $\bar{g} \cdot \tau = \tau$ with $\bar{g} \in PSL(2, \mathbb{Z})$ and $\bar{g} \neq 1$ then either

1. $\tau = i$ and the stabilizer group is $\{1, \bar{S}\}$
2. $\tau = \omega = e^{2\pi i/3} = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$ and the stabilizer group is

$$\{1, \bar{S}\bar{T}, (\bar{S}\bar{T})^2\} \quad (7.133)$$

3. $\tau = -\omega^2 = e^{\pi i/3} = \frac{1}{2} + \frac{\sqrt{3}}{2}i$ and the stabilizer group is

$$\{1, \bar{T}\bar{S}, (\bar{T}\bar{S})^2\} \quad (7.134)$$

In particular, for all other points $\tau \in \tilde{\mathcal{F}}$, the stabilizer group is the trivial group.

Lemma 2 follows quickly from Lemma 1: If $g\tau = \tau$ then we must be in the case $c = \pm 1$. If $d = 0$ then $a - 1/\tau = \tau$ for some integer a . But we must also have $|\tau| = 1$ and hence $a = \tau + \bar{\tau}$. Since a is an integer we quickly find that $a = 0$ with $\tau = i$, or $a = \pm 1$ with $\tau = \omega$ or $-\omega^2$. If $d = \pm 1$ then from the saturation condition $2cdx = -|cd|$ we get $x = -\frac{1}{2}d/|d|$ and hence $\tau = \omega$ or $-\omega^2$.

Now we can finally prove:

Theorem: $SL(2, \mathbb{Z})$ is generated by S and T .

Proof: Let τ_0 be in the interior of $\tilde{\mathcal{F}}$. Then choose any element $g \in SL(2, \mathbb{Z})$ with $\bar{g} \neq 1$. Then there is an element $g' \in \bar{G}$ so that $g'g \cdot \tau_0 \in \tilde{\mathcal{F}}$. Moreover, this element must be in the interior of $\tilde{\mathcal{F}}$ by Lemma 1 and hence, by Lemma 2, $g'g = 1$ in $PSL(2, \mathbb{Z})$. Therefore $\bar{g} \in \bar{G}$, which means $\bar{G} = PSL(2, \mathbb{Z})$. Moreover, $S^2 = -1$, and hence S and T generate $SL(2, \mathbb{Z})$. ♠

1. The exact fundamental domain $\mathcal{F}$ must be chosen so that no two distinct points on the boundary are G -related. So, for example, we could choose the part of the boundary with $\operatorname{Re}(\tau) \geq 0$.
2. One can show that the relations on S and T are:

$$S^2 = -1 \quad (ST)^3 = -1 \quad (7.135)$$

3. Given $g \in SL(2, \mathbb{Z})$ it is possible to write the word in S, T giving g by applying the Euclidean algorithm to (a, c) and interpreting the standard equations there in terms of matrices. See Chapter 1, Section 8.

4. Although the keyhole region is the standard fundamental domain there is no unique choice of fundamental domain. For example, one could equally well use any of the images shown in Figure 58 (and of course there are infinitely many such regions). Moreover, we could displace $\mathcal{F} \rightarrow \mathcal{F} + \epsilon$ and still produce a fundamental domain.
5. The action of $PSL(2, \mathbb{Z})$ is properly discontinuous on $\mathcal{H}$, but not quite free. If we consider finite-index subgroups that do not contain the stabilizer groups mentioned above then the action will be free and the quotient space will be a nice Riemann surface.

Exercise

- a.) The *face-centered-cubic* lattice in $\mathbb{R}^3$ is the sublattice of $\mathbb{Z}^3$ of all points such that $\sum_i x_i \equiv 0 \pmod{2}$. Construct the Wigner-Seitz cell for the fcc lattice.
 - b.) The *body-centered-cubic* lattice in $\mathbb{R}^3$ is the sublattice of $\mathbb{Z}^3$ such that the coordinates x_i are all even or all odd. Construct the Wigner-Seitz cell for the bcc lattice.
-

Exercise

- a.) The group $\Gamma_0(2)$ is the subgroup of $SL(2, \mathbb{Z})$ of matrices with $c \equiv 0 \pmod{2}$. Find a fundamental domain for $\bar{\Gamma}_0(2)$ acting on $\mathcal{H}$.
 - b.) The group $\Gamma(2)$ is the subgroup of $SL(2, \mathbb{Z})$ of matrices congruent to 1 modulo 2. Find a fundamental domain for $\bar{\Gamma}(2)$.
-

7.7.3 Algebras and double cosets

TO BE WRITTEN

7.7.4 Orbifolds

An interesting class of examples where the quotient space is “almost” a manifold are called “orbifolds” or “V-manifolds.”

In a manifold, neighborhoods of points locally look like copies of $\mathbb{R}^n$.

In an orbifold neighborhoods of points locally look like copies of $\mathbb{R}^n/\Gamma$ where Γ is a finite group acting on $\mathbb{R}^n$. The finite group Γ can depend on the point in question. For example, for “most” points it might be trivial. But for some other points it might be nontrivial

Example 1: Let $\mathbb{Z}_2$ act on $\mathbb{R}^n$ by $\sigma \cdot \vec{x} = -\vec{x}$. Then in the quotient space $\mathbb{R}^n/\mathbb{Z}_2$ the neighborhood of every point is homeomorphic to $\mathbb{R}^n$ except for the origin. A neighborhood of $[\vec{0}]$ is a cone on $\mathbb{RP}^{n-1}$. Since $\mathbb{RP}^{n-1}$ is not homotopy equivalent to S^{n-1} for $n > 2$ this space cannot be homeomorphic to $\mathbb{R}^n$ for $n > 2$.

Example 2: Consider the n -dimensional torus $T^n = \mathbb{R}^n/\mathbb{Z}^n$. Let $\Gamma = \mathbb{Z}_2$ act on T^n with the action induced from $\sigma \cdot \vec{x} = -\vec{x}$ on $\mathbb{R}^n$. There are now 2^n fixed points. At each fixed point the local neighborhood is homeomorphic to $\mathbb{R}^n/\mathbb{Z}_2$.

Example 3: Let $G \cong \mathbb{Z}/N\mathbb{Z}$ be the group of N^{th} roots of 1 acting on the complex plane by multiplication. The quotient $\mathbb{C}/\mathbb{Z}_N$ is an orbifold. The stabilizer group is trivial everywhere except the origin, where it is all of G . A neighborhood of 0 should be viewed as a cone with opening angle $2\pi/N$.

Example 4: Recall that we showed that we can identify $\mathcal{H} \cong SL(2, \mathbb{R})/SO(2)$. Now consider the double-coset

$$\Gamma \backslash SL(2, \mathbb{R})/SO(2) \quad (7.136)$$

where $\Gamma = SL(2, \mathbb{Z})$. We can identify this as the set of orbits of Γ on $\mathcal{H}$. From our analysis of the fundamental domain above we see that it is topologically a sphere with two orbifold singularities. The one at $[i]$ has a neighborhood modeled on $\mathbb{C}/\mathbb{Z}_2$ and the one at $[\omega] = [-\bar{\omega}]$ has a neighborhood modeled on $\mathbb{C}/\mathbb{Z}_3$. In addition, there is one puncture (at $\tau = i\infty$).

Exercise *Weighted projective spaces*

Choose positive integers $p_1, \dots, p_{n+1}$. Then the *weighted projective space* $W\mathbb{P}[p_1, \dots, p_{n+1}]$ is the space defined by $(\mathbb{C}^{n+1} - \{\vec{0}\})/\mathbb{C}^*$ with $\mathbb{C}^*$ action:

$$\lambda \cdot (z_1, \dots, z_{n+1}) := (\lambda^{p_1} z_1, \dots, \lambda^{p_{n+1}} z_{n+1}) \quad (7.137)$$

Show that the resulting quotient space is a well-defined orbifold. What are the orbifold singularities?

7.7.5 Examples of quotients which are not manifolds

When we divide by a *noncompact group* like $\mathbb{C}^*$ or $GL(n, \mathbb{C})$ then, even when the action is free the quotient space M/G can be a “bad” (e.g. non-Hausdorff, or very singular and difficult-to-work-with) space.

Example 1 Consider the circle $S^1 \cong \mathbb{R}/\mathbb{Z}$. Consider translation $x \rightarrow x + \alpha$ where α is *irrational*. This generates an action of $\mathbb{Z}$ on S^1 (or $\mathbb{Z} \oplus \mathbb{Z}$ on $\mathbb{R}$) that is *not* properly discontinuous. The quotient space is *not* a manifold.

Example 2 . A similar example. Consider the two-dimensional torus, identified with $S^1 \times S^1 \cong \mathbb{R}/\mathbb{Z} \times \mathbb{R}/\mathbb{Z}$. Thus we can use coordinates (x, y) where x, y are defined modulo 1. Consider the group action by $\mathbb{R}$:

$$(x, y) \rightarrow (x + tv_1, y + tv_2) \quad (7.138)$$

where $t \in \mathbb{R}$, (v_1, v_2) is some vector. If the slope v_2/v_1 is a rational number the orbits are compact and the space of orbits is a nice space. (Exercise: What is this space?) But if the

slope is irrational we again have a non-Hausdorff space since the orbits are all dense, so it is impossible to separate open sets around these orbits.

Example 3: Consider $X = \mathbb{C}^{n+1}$ and $G = \mathbb{C}^*$ acting as

$$\lambda \cdot \vec{z} = \lambda \cdot (z_1, \dots, z_{n+1}) := (\lambda z_1, \dots, \lambda z_{n+1}) \quad (7.139)$$

for $\lambda \in \mathbb{C}^*$. Then $X/\mathbb{C}^*$ is not Hausdorff. For consider the equivalence class $[\vec{0}]$. Let $\pi : X \rightarrow X/\mathbb{C}^*$ be the projection. An open neighborhood U of $[\vec{0}]$ is such that $\pi^{-1}(U)$ is an open neighborhood of $\vec{0}$. Now consider any other point $[\vec{w}]$ where $\vec{w} \neq 0$. There will always be a $\lambda \in \mathbb{C}^*$ so that $\lambda \vec{w} \in \pi^{-1}(U)$. Therefore, $[\vec{w}] \in U$ for *all* points $[\vec{w}] \in X/\mathbb{C}^*$! In particular one cannot separate $[\vec{0}]$ and $[\vec{w}]$ by open sets. Clearly, there is one bad actor here, the point $\vec{0}$. Indeed if we eliminate it then the $\mathbb{C}^*$ action on $\mathbb{C}^* - \{0\}$ produces a good manifold

$$\mathbb{C}P^n = (\mathbb{C}^* - \{0\})/\mathbb{C}^* \quad (7.140)$$

Example 4: For a very similar example let us consider a $\mathbb{C}^*$ action on $\mathbb{C}^2$ with two coordinates ϕ_1, ϕ_2 and action:

$$\begin{aligned} \phi_1 &\rightarrow \lambda \phi_1 \\ \phi_2 &\rightarrow \lambda^{-1} \phi_2 \end{aligned} \quad (7.141)$$

Generic orbits are labelled by the “gauge invariant” quantity $\phi_1 \phi_2$. Note however that there are 3 special orbits corresponding to the value $\phi_1 \phi_2 = 0$:

$$\begin{aligned} \mathcal{O}_1 &= \{(\phi_1, 0) | \phi_1 \neq 0\} \\ \mathcal{O}_2 &= \{(0, \phi_2) | \phi_2 \neq 0\} \\ \mathcal{O}_3 &= \{(0, 0)\} \end{aligned} \quad (7.142)$$

Unlike projective space now the space

$$[\mathbb{C}^2 - \mathcal{O}_3]/\mathbb{C}^* \quad (7.143)$$

is NOT a Hausdorff space. Indeed the orbits $\mathcal{O}_1$ and $\mathcal{O}_2$ project to two distinct points $[\mathcal{O}_1]$ and $[\mathcal{O}_2]$ and we claim these cannot be separated from each other by open sets in the quotient topology. See the exercise below.

We could just consider the space

$$[\mathbb{C}^2 - (\mathcal{O}_1 \cup \mathcal{O}_2 \cup \mathcal{O}_3)]/\mathbb{C}^* \quad (7.144)$$

Now the quotient by $\mathbb{C}^*$ is Hausdorff and is a nice algebraic variety. It is just a copy of $\mathbb{C}^*$.

But we could compactify (7.144). One way to do this is to omit the axes $\phi_1 = 0$ and $\phi_2 = 0$, but include the origin, then the quotient space $\mathbb{C}^2 - (\mathcal{O}_1 \cup \mathcal{O}_2)$ is well-defined and

in fact isomorphic to $\mathbb{C}$. But in fact, there are three ways of getting a good quotient by omitting any two of the three special orbits and then dividing by $\mathbb{C}^*$.

Remark: We did not use the complex numbers in an essential way in Example 4 so we can apply the very same ideas to the quotient of Minkowski space by the component of the identity of the Lorentz group. We can consider the real subspace $\mathbb{R}^2 \subset \mathbb{C}^2$ and interpret ϕ_1, ϕ_2 as light-cone coordinates. Then we consider the subgroup of $\mathbb{C}^*$ corresponding to $\mathbb{R}_+$. We studied the orbits in Section **** above. Using the same kind of reasoning as above we see that the quotient is not a Hausdorff space since the origin cannot be separated from the lightlike orbits. Even if we remove the origin the quotient is not Hausdorff since we cannot separate timelike and spacelike orbits from lightlike orbits. If we remove the origin and the lightlike orbits *then* the quotient space is a Hausdorff space. It is naturally thought of as four copies of $\mathbb{R}_+$.

Example 5: Another example similar to the above is the following. Let us consider the set of conjugacy classes of all matrices in $M \in M_n(\mathbb{C})$ under the action of $S \in GL(n, \mathbb{C})$:

$$M \mapsto SMS^{-1} \quad (7.145)$$

Let us consider the conjugacy classes of matrices:

$$\mathcal{C}(n) := M_n(\mathbb{C})/GL(n, \mathbb{C}) \quad (7.146)$$

we could consider this as the set of linear transformations $T : V \rightarrow V$ up to change of basis, where V is an n -dimensional complex vector space.

The key fact here is that every matrix can be brought to Jordan canonical form: (See Linear Algebra User's Manual, chapter **** for more details, application, and a proof.)

Take the characteristic polynomial $p(x) = \det(x\mathbf{1} - M)$ and factor it into its distinct complex roots $p(x) = \prod_{i=1}^s (x - \lambda_i)^{r_i}$ where $r_i > 0$. Then, first of all we can bring M to block diagonal form $M = \oplus_{i=1}^s M_i$, where each block M_i is itself a block of Jordan matrices of type i , that is, there are positive integers n_α^i , $1 \leq \alpha \leq k_i$ with $\sum_\alpha n_\alpha^i = k_i$ such that

$$M_i = \oplus_{\alpha=1}^{k_i} J_{\lambda_i}^{(n_\alpha^i)} \quad (7.147)$$

where $J_\lambda^{(1)}$ is the 1×1 matrix with entry λ and, for $n > 1$, $J_\lambda^{(n)} = \lambda \mathbf{1}_{n \times n} + \sum_{j=1}^{n-1} e_{j,j+1}$ cannot be brought to a simpler form, and certainly cannot be diagonalized.

If a matrix M has nontrivial Jordan form then the quotient space $M_n(\mathbb{C})/GL(n, \mathbb{C})$ is non-Hausdorff in the neighborhood of $[M]$. To see why it suffices to consider the case $n = 2$: The characteristic polynomial $p(x) := \det(x\mathbf{1} - M)$ is now order two. When there are two distinct roots M is diagonalizable. When two roots coincide, say $p(x) = (x - \lambda)^2$, then there are two possibilities: Either $M = \lambda \mathbf{1}_{2 \times 2}$ for some λ , or, M is conjugate to Jordan form:

$$M = S \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} S^{-1} \quad (7.148)$$

The existence of matrices with nontrivial Jordan blocks leads to non-Hausdorff behavior of the quotient. To see this it suffices to consider conjugation by matrices of the form:

$$g = \begin{pmatrix} t_1 & 0 \\ 0 & t_2 \end{pmatrix} \in GL(2, \mathbb{C}) \quad (7.149)$$

then

$$g \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} g^{-1} = \begin{pmatrix} \lambda & z \\ 0 & \lambda \end{pmatrix} \quad (7.150)$$

with $z = t_1/t_2$ so

$$\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \sim \begin{pmatrix} \lambda & z \\ 0 & \lambda \end{pmatrix} \quad (7.151)$$

for any $z \in \mathbb{C}^*$.

Now, suppose $\bar{V} \subset \mathcal{C}(2)$ is an open set in the quotient topology and moreover

$$\left[\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \right] \in \bar{V} \quad (7.152)$$

$p^{-1}(\bar{V})$ is an open set in $M_2(\mathbb{C})$ containing the matrix $\lambda \mathbf{1}_{2 \times 2}$. But any such open set must also contain

$$\begin{pmatrix} \lambda & z \\ 0 & \lambda \end{pmatrix} \quad (7.153)$$

for some sufficiently small z . Therefore,

$$\left[\begin{pmatrix} \lambda & z \\ 0 & \lambda \end{pmatrix} \right] = \left[\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \right] \quad (7.154)$$

for $z \neq 0$, is a distinct point from

$$\left[\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \right] \quad (7.155)$$

and yet any neighborhood of $\left[\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \right]$ contains $\left[\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \right]$. So, these distinct points cannot be separated by open sets.

If, by hand, we consider the subset of diagonalizable matrices in $M_2(\mathbb{C})$ then the quotient is better. If $m \in GL(2, \mathbb{C})$ is of the form:

$$m = g \begin{pmatrix} z_1 & 0 \\ 0 & z_2 \end{pmatrix} g^{-1} \quad z_1 \neq z_2 \quad (7.156)$$

then $[m]$ is parametrized by the *unordered* pair $\{z_1, z_2\}$. This is just the configuration space $\mathbb{C}^2/\mathbb{Z}_2$ and is just an orbifold, with a $\mathbb{Z}_2$ orbifold singularity along the diagonal $\{z, z\}$.

Returning to the general case, if we consider $M_n^{\text{ss}}(\mathbb{C}) \subset M_n(\mathbb{C})$ of diagonalizable matrices then the quotient $M_n^{\text{ss}}(\mathbb{C})/GL(n, \mathbb{C})$ is much better behaved. It is an orbifold, with

orbifold singularities corresponding to various symmetric groups along loci where eigenvalues coincide.

Remarks

1. Making nice quotients is part of the general subject of “geometric invariant theory.” See, Fogarty, Kirwan, and Mumford, *Geometric Invariant Theory* for a sophisticated presentation of conditions in algebraic geometry for when quotients are “good.”
2. The mathematics of Geometric Invariant theory finds several applications in supersymmetric field theory and string theory. First of all, it is important in forming various moduli spaces. (For example, we saw it was necessary in forming a good quotient space corresponding to the moduli space of lines.) Closely related to this, it is important in understanding the moduli spaces of vacua in SUSY field theory.
3. The examples of quotients by ergodic actions such as the irrational rotation on a circle forms one of the primary examples in the study of *noncommutative geometry*. See, A. Connes, *Noncommutative Geometry* for much detailed discussion.

Exercise

Consider a small open neighborhood of $[\mathcal{O}_1]$:

$$\{[(1, z_2)] : |z_2| < \epsilon\} \quad (7.157)$$

and a small open neighborhood of $[\mathcal{O}_2]$:

$$\{[(z_1, 1)] : |z_1| < \epsilon'\} \quad (7.158)$$

a.) Show that when $n > 0$ in (7.141) these sets will intersect each other no matter how small we take ϵ, ϵ'

Therefore we cannot separate $[\mathcal{O}_1]$ from $[\mathcal{O}_2]$ and the quotient space is not Hausdorff.

7.7.6 When is the quotient of a manifold by an equivalence relation another manifold?

A natural question which arises from these examples is, more generally, when the quotient of a manifold by a general equivalence relation is another manifold.

An equivalence relation $\sim$ on any set X has a *graph* which is the subset $R \subset X \times X$ defined by

$$R = \{(x, y) | x \sim y\} \quad (7.159)$$

Conversely, one could define an equivalence relation from subsets of $X \times X$ satisfying certain properties.

We have already discussed how, if X is a topological space and $\sim$ is an equivalence relation then we can define a topological space with a continuous projection $p : X \rightarrow X/\sim$. A question which frequently arises is this: If M is a manifold, when is $M/\sim$ a manifold?

One criterion for answering this question is the following theorem: ⁶¹

Theorem: If M is a smooth manifold and $\sim$ is an equivalence relation then the following are equivalent:

1. $M/\sim$ is a smooth manifold and $p : M \rightarrow M/\sim$ is a submersion.
2. The graph $R \subset M \times M$ is a proper smooth submanifold and the projection $p_1 : M \times M$ onto the first factor, when restricted to R , is a proper submersion.

8. Homogeneous spaces of Lie groups

If H is a subgroup of G then we have defined above a *homogeneous space* as the set G/H , the set of right H -orbits on G , or the set of left H -cosets.

When G is a Lie group and H is a Lie subgroup then, when the conditions of the quotient manifold theorem apply, the homogeneous space G/H is a beautiful manifold with a high degree of symmetry. Note that the right-action of H on G is clearly smooth and free, so the only question is whether it is proper. If G and H are compact, this is automatic. When G and/or H is noncompact it might or might not be true. When the quotient manifold theorem applies the dimension of G/H is $\dim G - \dim H$.

In fact, more is true: Recall that $T_1G = \mathfrak{g}$ is the Lie algebra of G and $T_1H = \mathfrak{h}$ is the Lie subalgebra of H . The tangent space is naturally identified with the quotient $\mathfrak{g}/\mathfrak{h}$. (This is not a Lie algebra, in general.) If we use a metric to define an orthogonal complement $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{p}$ then we can identify $\mathfrak{p} \cong T_1(G/H)$.

One can usefully introduce metrics into the discussion. For matrix groups we can define a metric on $\mathfrak{g}$ by $(X, Y) = -\text{Tr}_V \rho(X)\rho(Y)$ where the trace is in some representation. If G is simple and connected then the metric only depends on the choice of V up to an overall constant and is proportional to the *Cartan-Killing form*:

$$(X, Y)_{C.K.} := -\text{Tr}_{\mathfrak{g}} \text{Ad}(X)\text{Ad}(Y) \quad (8.1)$$

In this case $\mathfrak{p}$ inherits a metric and hence G/H inherits a metric which can be pushed forward globally over G/H .

Example 1: Let us consider $SO(3, \mathbb{R})/SO(2, \mathbb{R})$ (where we choose a particular subgroup of $SO(3, \mathbb{R})$ isomorphic to $SO(2, \mathbb{R})$.) We have already seen from the stabilizer-orbit theorem that as a set:

$$SO(3, \mathbb{R})/SO(2, \mathbb{R}) \longleftrightarrow S^2 \quad (8.2)$$

⁶¹For a proof see Theorem 8.3 of the notes “Differential Geometry,” by R.L. Fernandes, in <http://www.math.illinois.edu/~ruiloja/Math519/notes.pdf>.

Now, from the quotient manifold theorem we can interpret this as a statement about smooth manifolds. In fact, we can even interpret it as a statement about Riemannian manifolds, when we use a Cartan-Killing metric on $SO(3, \mathbb{R})$.

Example 2: Generalizing the previous example, consider $SO(n+1)/SO(n)$. Note that $SO(n+1)$ acts transitively on the unit n -sphere in $\mathbb{R}^{n+1}$ with stabilizer $SO(n)$. For example, choose $\vec{x}_0 = (0, \dots, 0, 1)$ then the stabilizer group is the $SO(n)$ subgroup

$$\begin{pmatrix} g & 0 \\ 0 & 1 \end{pmatrix} \quad (8.3)$$

with $g \in SO(n)$. Therefore $S^n = SO(n+1)/SO(n)$ as manifolds. Recall that the Lie algebra of $SO(n+1)$ is the set of real antisymmetric matrices. Using the inner product

$$(X, Y) := -\text{Tr}_{n+1} XY \quad (8.4)$$

we can take $\mathfrak{p}$ to be the span of the antisymmetric matrices $p_i = e_{i,n+1} - e_{n+1,i}$, $1 \leq i \leq n$. These represent tangent vectors on the sphere and the metric is $(p_i, p_j) = 2\delta_{i,j}$.

Example 3: As we have seen, the group $SL(2, \mathbb{R})$ has a natural left-action on the upper-half-plane via fractional linear transformations:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \tau := \frac{a\tau + b}{c\tau + d} \quad (8.5)$$

It is not difficult to show that this is a *transitive* action on the upper-half-plane. It is also particularly easy to compute the stabilizer subgroup of $\tau = i$. Converting the left-action to a right-action (as is often convenient in applications) via $\tau \mapsto g^{-1} \cdot \tau$ we can identify:

$$\mathcal{H} \cong SL(2, \mathbb{R})/SO(2) \quad (8.6)$$

In this case, the natural metric is an extremely interesting metric known as the *Poincaré metric*

$$ds^2 = \frac{|d\tau|^2}{(\text{Im}\tau)^2} \quad (8.7)$$

In physics it is the first example of a Euclidean anti-deSitter metric.

Remarks

1. A common application of homogeneous spaces is in describing vacua of a scalar field theory with a global symmetry. Suppose ϕ is a scalar field on d -dimensional Minkowski space, $\mathbb{M}^{1,d-1}$, valued in some representation space V of a compact Lie group G . Suppose that $U(\phi)$ is a G -invariant potential energy. Typically it is an

invariant polynomial in ϕ which is bounded below for a stable and renormalizable field theory. If we choose a G -invariant metric on V then the action will be

$$S[\phi] = \int_{\mathbb{M}^{1,d-1}} \left[\frac{1}{2}(\partial_\mu \phi, \partial^\mu \phi) - U(\phi) \right] d^d x \quad (8.8)$$

so the energy will be

$$H[\phi] = \int_{\mathbb{R}^{d-1}} \left[\frac{1}{2}(\Pi, \Pi) + \frac{1}{2}(\partial_i \phi, \partial_i \phi) + U(\phi) \right] d^d x \quad (8.9)$$

Then, in classical field theory, the vacua are field configurations where the momentum $\Pi = 0$, the field is constant $\partial_i \phi = 0$ and the constant value of ϕ minimizes $U(\phi)$. Of course if $\phi \in V$ minimizes U then so does $\phi(g)\phi$ for any $g \in G$. It follows that the set of classical vacua, $\mathcal{M}$, has an action of G on it, and therefore decomposes as a union of G -orbits.

Example 1: Suppose $\phi \in \mathbb{R}^n$ is a real-valued vector and $G = SO(n)$. Then if

$$U(\phi) = \lambda(\phi \cdot \phi - v^2)^2, \quad (8.10)$$

where the coupling constant $\lambda > 0$ so that the energy is bounded below, the set of minima form the sphere $\mathcal{M} \cong S^{n-1} \cong SO(n)/SO(n-1)$ of radius $|v|$.

Example 2: Suppose $\phi \in \mathcal{H}_n$ is a Hermitian $n \times n$ matrix and $G = SU(n)$ acts on $\mathcal{H}_n$ by $\rho(u) \cdot \phi = u\phi u^\dagger$. Suppose that

$$U(\phi) = \lambda \text{Tr}(\phi^2 - v^2)^2, \quad (8.11)$$

where again $\lambda > 0$ is a coupling constant. Since $(\phi^2 - v^2)^2$ is a positive definite Hermitian matrix the minimum is obtained by $(\phi^2 - v^2)^2 = 0$. Again, using the fact that ϕ is Hermitian it follows that $\phi^2 - v^2 = 0$. Any such matrix has the property that

$$P = \frac{1}{2}(1 + v^{-1}\phi) \quad (8.12)$$

is an Hermitian projection operator. The image of P is a k -dimensional subspace of $\mathbb{C}^n$. The set of classical vacua is isomorphic to the disjoint union of Grassmannians:

$$\mathcal{M} \cong \coprod_{k=0}^n \text{Gr}(k, n) \quad (8.13)$$

In quantum field theory with $d > 2$ the components $\text{Gr}(k, n)$, if not lifted by quantum effects, will correspond to a family of vacua in which $SU(n)$ is spontaneously broken to a subgroup isomorphic to $S(U(k) \times U(n-k))$.

Exercise

Let $\pi \in M_{n_1 \times n_2}(\mathbb{C})$ be an $n_1 \times n_2$ complex matrix.

a.) Show that

$$\exp \begin{pmatrix} 0 & \pi \\ -\pi^\dagger & 0 \end{pmatrix} = \begin{pmatrix} \cos \left(\sqrt{\pi \pi^\dagger} \right) & \pi \frac{\sin(\sqrt{\pi^\dagger \pi})}{\sqrt{\pi^\dagger \pi}} \\ -\pi^\dagger \frac{\sin(\sqrt{\pi \pi^\dagger})}{\sqrt{\pi \pi^\dagger}} & \cos \left(\sqrt{\pi^\dagger \pi} \right) \end{pmatrix} \quad (8.14)$$

b.) In particular, show that for z a complex number we have

$$\exp \begin{pmatrix} 0 & -\bar{z} \\ z & 0 \end{pmatrix} = \frac{1}{\sqrt{1 + |u|^2}} \begin{pmatrix} 1 & -\bar{u} \\ u & 1 \end{pmatrix} \quad (8.15)$$

with

$$u = z \frac{\tan |z|}{|z|} \quad (8.16)$$

c.) Show that $SU(2)/U(1) \cong \mathbb{CP}^1$ by considering the linear action of $SU(2)$ on homogeneous coordinates of $\mathbb{CP}^1$. Using this action show that we can interpret the parameter u in part (b) as the stereographic projection of the sphere to the complex plane.

d.) Show that for $G = SO(n+1)$ taking $\pi = \sum_{i=1}^n \theta^i p_i$ we have

$$\exp[\pi] \vec{x}_0 = \begin{pmatrix} \frac{\sin |\theta|}{|\theta|} \theta^i \\ \cos |\theta| \end{pmatrix} \quad (8.17)$$

where $p_i = e_{i,n+1} - e_{n+1,i}$, $|\theta| = \sqrt{\sum (\theta^i)^2}$ and $\vec{x}_0$ has all components zero except for the last one, which is one.

e.) Show that

$$\exp \begin{pmatrix} 0 & \pi \\ \pi^\dagger & 0 \end{pmatrix} = \begin{pmatrix} \cosh \left(\sqrt{\pi \pi^\dagger} \right) & \pi \frac{\sinh(\sqrt{\pi^\dagger \pi})}{\sqrt{\pi^\dagger \pi}} \\ \pi^\dagger \frac{\sinh(\sqrt{\pi \pi^\dagger})}{\sqrt{\pi \pi^\dagger}} & \cosh \left(\sqrt{\pi^\dagger \pi} \right) \end{pmatrix} \quad (8.18)$$

Remark: Note well that (8.14) forms a compact set of matrices and (8.18) forms a noncompact set of matrices.

f.) Consider the Lorentz group on $\mathbb{M}^{1,d-1}$. Using (e) compute the Lorentz matrix corresponding to a boost of rapidity β in the $\hat{n}$ direction.

Exercise

a.) Show that $O(n+1)/O(n) \cong S^n$.

b.) Show that $O(n+1)/SO(n) \cong S^n \amalg S^n$.⁶²

⁶²Hint: Consider a transitive action of $O(n+1)$ on $S^n \times \mathbb{Z}_2$ and compute the stabilizer.

Exercise

- a.) Show that $SU(n+1)/SU(n) \cong S^{2n-1}$.
 - b.) Show that $SU(n+1)/(SU(n) \times U(1)) \cong \mathbb{CP}^{n-1}$.
 - c.) Show that there is an action of $U(1)$ on S^{2n-1} so that $S^{2n-1}/U(1) \cong \mathbb{CP}^{n-1}$.
-

8.1 Grassmannians

Let $\mathbb{F}$ stand for $\mathbb{R}, \mathbb{C}$, or $\mathbb{H}$. The Grassmannian $Gr_{k,n}$, or sometimes $Gr_k(\mathbb{F}^n)$, is the moduli space of k -planes in $\mathbb{F}^n$.

8.1.1 Homogeneous spaces

We will first describe the Grassmannian as a homogeneous space in several different ways.

First, note that $GL(n)$ (meaning $GL(n, \mathbb{R})$ or $GL(n, \mathbb{C})$ for $\mathbb{F} = \mathbb{R}, \mathbb{C}$) acts on the Grassmannian: If $V \subset \mathbb{F}^n$ is a k -dimensional subspace, its image under an invertible linear transformation is another k -dimensional subspace.

We claim the action is transitive: Choose a standard basis $e_1, \dots, e_n$ for $\mathbb{F}^n$. Let V_0 be the span of $e_1, \dots, e_k$. Suppose V is any other subspace, and choose some basis $w_1, \dots, w_k$ for V . There is certainly a linear transformation taking $e_1, \dots, e_k$ to $w_1, \dots, w_k$. By completing the basis in different ways we can complete the definition of the relevant linear transformation. In general $g \in GL(n, \mathbb{F})$ acts by

$$g \cdot e_i = \sum_{j=1}^n g_{ji} e_j \quad (8.19)$$

so the stabilizer of V_0 is the set of invertible matrices of the form

$$\begin{pmatrix} A & B \\ 0 & D \end{pmatrix} \quad (8.20)$$

where A is an invertible $k \times k$ matrix and D is an invertible $(n-k) \times (n-k)$ matrix. The group of these matrices, denoted P is an example of a “maximal parabolic subgroup.” Replacing g by g^{-1} so that our action is a right action we can say that

$$Gr_{k,n} \cong GL(n)/P \quad (8.21)$$

Since the groups involved are noncompact it is not quite obvious from the stabilizer orbit theorem that the quotient is a smooth manifold. Other representations below will make it obvious that it is. We can therefore say that the dimension is $\dim GL(n) - \dim P$ and this is just the dimension of the southwest block in (8.20). Therefore the dimension is just $k(n-k)$. But now we need to be careful about coefficients:

$$\begin{aligned} \dim_{\mathbb{R}} Gr_{k,n} &= k(n-k) & \mathbb{F} &= \mathbb{R} \\ \dim_{\mathbb{R}} Gr_{k,n} &= 2k(n-k) & \mathbb{F} &= \mathbb{C} \\ \dim_{\mathbb{R}} Gr_{k,n} &= 4k(n-k) & \mathbb{F} &= \mathbb{H} \end{aligned} \quad (8.22)$$

There is another, very useful, way to look at the Grassmannian as a homogeneous space:

We introduce

$$V_k^{nc}(\mathbb{F}^n) \subset \text{Mat}_{k \times n}(\mathbb{F}) \quad (8.23)$$

where $V_k^{nc}(\mathbb{F}^n)$ is the set of $k \times n$ matrices of rank k . We denote a typical element as Λ and write:

$$\Lambda = \begin{bmatrix} v_{11} & \cdots & \cdots & \cdots & \cdots & v_{1n} \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ v_{k1} & \cdots & \cdots & \cdots & \cdots & v_{kn} \end{bmatrix} \quad (8.24)$$

On the one hand, to any such matrix Λ we can associate a k -dimensional subspace of $\mathbb{F}^n$: We simply regard the rows as defining k different vectors in $\mathbb{F}^n$ and we let $V(\Lambda)$ be the subspace spanned by these vectors. Since Λ has rank k the subspace $V(\Lambda)$ is of dimension k (over $\mathbb{F}$). Thus $V_k^{nc}(\mathbb{F}^n)$ is the moduli space of k -dimensional subspaces in $\mathbb{F}^n$ equipped with a basis. A change of bases corresponds to multiplication of Λ on the left by $h \in GL(k)$. Thus $V(h\Lambda) = V(\Lambda)$. Conversely, given any k -dimensional subspace $V \subset \mathbb{F}^n$ we can choose any basis for V and form the corresponding matrix Λ . Thus, we can identify:

$$Gr_{k,n} = GL(k) \backslash V_k^{nc}(\mathbb{F}^n) \quad (8.25)$$

We denote equivalence classes by $[\Lambda]$.

Now, let us notice that there is also a right action of $GL(n)$ on $Gr_{k,n}$ which is obvious in the above description. The action

$$\Lambda \rightarrow \Lambda \cdot g \quad (8.26)$$

preserves $V_k^{nc}(\mathbb{F}^n)$ since it cannot change the rank. Moreover this $GL(n)$ action commutes with the left action of $GL(k)$, so it is a well-defined action on the $GL(k)$ equivalence classes $[\Lambda]g = [\Lambda \cdot g]$.

We have already seen that the $GL(n)$ action is transitive on $Gr_{k,n}$. By changing $h \rightarrow h^{-1}$, (or $g \rightarrow g^{-1}$) we obtain a right or left $GL(k) \times GL(n)$ action on $V_k^{nc}(\mathbb{F}^n)$. This action is *transitive*. This is simply the fact that under the arbitrary basis change of domain and range of a linear transformation the only invariant is the rank. In terms of matrices we consider the right-action by $GL(k) \times GL(n)$ on $V_k^{nc}(\mathbb{F}^n)$:

$$\Lambda \rightarrow h^{-1} \Lambda g \quad (8.27)$$

Then we can always bring Λ to the form:

$$\begin{pmatrix} \mathbf{1}_{k \times k} & \mathbf{0}_{k \times (n-k)} \end{pmatrix} \quad (8.28)$$

The stabilizer of this point in $V_k^{nc}(\mathbb{F}^n)$ is the subgroup of $GL(k) \times GL(n)$ of matrices of the form:

$$\left(A^{-1}, \begin{pmatrix} A & 0 \\ C & D \end{pmatrix} \right) \quad (8.29)$$

The group of these matrices is isomorphic to P . Thus

$$V_k^{nc}(\mathbb{F}^n) \cong GL(k) \times GL(n)/P \quad (8.30)$$

When we further mod out by $GL(k)$ to get $Gr_{k,n}$ we get (8.21), as advertised.

Remarks:

1. Let $T \subset GL(n)$ be the subgroup of diagonal invertible matrices. This group is isomorphic to $(\mathbb{F}^*)^n$. It acts on the Grassmannian. What is the set of fixed points? It is easy to see that, in the standard basis the fixed points are the subspaces generated by vectors $\{e_{j_1}, e_{j_2}, \dots, e_{j_k}\}$ where, WLOG, we can take:

$$1 \leq j_1 < j_2 < \dots < j_k \leq n. \quad (8.31)$$

We will denote such an increasing k -tuple as J . It is known as a *Schubert symbol*. The corresponding subspace will be V_J . So what we denoted V_0 above will now be denoted $V_{1,\dots,k}$. It is not difficult to show that these are the only fixed points. There are thus $\binom{n}{k}$ isolated fixed points of the torus action.

2. Note that for each Schubert symbol there is a corresponding $k \times k$ minor of Λ , which we can denote Λ^J . Let $P_J := \det \Lambda^J$. Then, since at least one minor is nonzero we can map Λ to a point in the projective space

$$[\dots : P^J : \dots] \in \mathbb{FP}^{N-1} \quad (8.32)$$

where $N = \binom{n}{k}$ and we choose, for example, lexicographic ordering of the J 's. Moreover, note that under $\Lambda \rightarrow h\Lambda$ the *Plücker coordinates* P^J change by $P^J \rightarrow (\det h)P^J$ and hence the map descends to a well-defined map $[\Lambda] \rightarrow [\dots : P^J : \dots]$. Thus, we have a map

$$Gr_{k,n} \rightarrow \mathbb{FP}^{N-1} \quad (8.33)$$

This turns out to be an embedding, known as the *Plücker embedding*. The image is characterized by a set of quadratic equations on the coordinates known as the *Plücker relations*.

8.1.2 Coordinates and coordinate patches

We now write down an atlas for the Grassmannian using the matrices Λ above.

Choose a standard basis $\vec{e}_1, \dots, \vec{e}_n$ for $\mathbb{F}^n$. We first describe a neighborhood of the point $V_{1\dots k}$. In this case the first k columns give a $k \times k$ invertible matrix and hence we can multiply by $GL(k)$ on the left to obtain a representative matrix in V_k^{nc} of the form:

$$\Lambda_0 = \begin{pmatrix} \mathbf{1}_{k \times k} & 0_{k \times (n-k)} \end{pmatrix} \quad (8.34)$$

Vector spaces in the neighborhood of V_0 will have Λ^{J_0} invertible where $J_0 = \{1, 2, \dots, k\}$ and hence can be brought to the form:

$$\Lambda = \begin{pmatrix} & a_{11} & \dots & a_{1,n-k} \\ \mathbf{1}_{k \times k} & \dots & \dots & \dots \\ & a_{k1} & \dots & a_{k,n-k} \end{pmatrix} \in Mat_{k \times n}(\mathbb{F}) \quad (8.35)$$

The coordinates are the numbers $a_{ij} \in \mathbb{F}$, in accord with the formula (8.22) for the dimensions. (Note, incidentally, that this identifies the tangent space with $\text{Hom}(V, \mathbb{F}^n/V)$.)

Why aren't the coordinates (8.35) global coordinates? The subspaces described by (8.35) have the property that if we let these vectors span a space W and we let V_c be the span of the *complementary* set of vectors $\vec{e}_{k+1}, \dots, \vec{e}_n$ then we have

$$W \cap V_c = \{0\} \quad (8.36)$$

for, if there were a nonzero vector in this intersection then its first k coordinates would be zero (since it is in V_c but then if it were in W it would have to vanish. On the other hand, there are clearly going to be k -dimensional subspaces $W \subset \mathbb{F}^n$ such that $W \cap V_c$ has positive dimension.

Choice of coordinate patches

The coordinate patches will now be labeled by Schubert symbols

$$J = \{1 \leq j_1 < j_2 < \dots < j_k \leq n\}. \quad (8.37)$$

We then define the corresponding coordinate patch to be:

$$U_J := \{W : \det \Lambda^J \neq 0\} \quad (8.38)$$

Note that it doesn't matter which representative Λ of the subspace W we choose. An equivalent (and more invariant) way to define U_J is to define the complementary subset $J_c = \{1, \dots, n\} - J$, introduce the notation:

$$V_{J,c} := \text{Span}\{\vec{e}_j\}_{j \in J_c} \quad (8.39)$$

and define the coordinate patch associated to J to be:

$$U_J \equiv \{W : W \cap V_{J,c} = \{\vec{0}\}\} \quad (8.40)$$

Our first job in checking this is an atlas is to show that it covers $Gr_{k,n}$. Suppose V is any subspace of $\mathbb{F}^n$ and choose a basis and hence a representative Λ . Some Plucker coordinate P_J must be nonzero, and hence some $k \times k$ minor must be nonzero. Then $V \in U_J$ for any such Schubert symbol.

To define coordinates on U_J we can choose a representative Λ by multiplying on the left by $g \in GL(k)$ so that $\Lambda^J = \mathbf{1}_{k \times k}$. We denote the resulting matrix by

$$\Lambda = (: 1_J \ A_J :) \quad (8.41)$$

where $A_J \in \text{Mat}_{k \times (n-k)}(\mathbb{F})$ and the $::$ sign means that the columns should be shuffled, without changing order, so that they are in the correct position.

On the overlap $U_J \cap U_{J'}$ of two patches we can choose a representative Λ for $W \in U_J \cap U_{J'}$ and $g_J, g_{J'}$ so that

$$\begin{aligned} g_J \Lambda &= (: 1_J \ A_J :) \\ g_{J'} \Lambda &= (: 1_{J'} \ A_{J'} :) \end{aligned} \quad (8.42)$$

and the transition function is given by

$$(: 1_J \ A_J :) = g_{J,J'} (: 1_{J'} \ A_{J'} :) \quad (8.43)$$

with

$$g_{J,J'} = g_J g_{J'}^{-1} \quad (8.44)$$

Note that we can write the matrix elements of $g_{J,J'}$ in terms of the matrix elements of A'_J or A_J and that the transition functions are thereby holomorphic for $\mathbb{F} = \mathbb{C}$. Thus, $Gr_{k,n}(\mathbb{C})$ is a complex manifold.

Exercise

Show that the coordinate atlas and transition functions for $\mathbb{F}\mathbb{P}^{n-1}$ are a special case of the above atlas in the case $Gr_{1,n}$.

8.1.3 Orthonormal bases

There is also a very useful viewpoint on Grassmannians which makes clear that they are compact. Let us put an hermitian metric on $\mathbb{F}^n$:

$$\| \vec{z} \|^2 := \sum_{i=1}^n z_i \bar{z}_i \quad (8.45)$$

Then we can always speak of orthonormal bases.

Definition. The Stiefel manifold $V_k(\mathbb{F}^n)$ is the space of matrices (8.24) such that the vectors are orthonormal. That is, it is the moduli space of k -planes equipped with an orthonormal basis.

Any two orthonormal bases of a vector space are related by $\Lambda \rightarrow g\Lambda$ with $g \in U_{\mathbb{F}}(k) = O(k), U(k), Sp(2k)$. In a manner directly analogous to the above we can write

$$V_k(\mathbb{F}^n) \cong U_{\mathbb{F}}(n)/U_{\mathbb{F}}(n-k) \quad (8.46)$$

$$Gr_{k,n} \cong U_{\mathbb{F}}(n)/U_{\mathbb{F}}(k) \times U_{\mathbb{F}}(n-k) \quad (8.47)$$

In homogeneous space coordinates, the patch around V_0 is described in terms of a complex $k \times (n-k)$ matrix π as $[\Lambda]$ where

$$\Lambda = \begin{bmatrix} \cos \sqrt{\pi \pi^\dagger} & \pi \frac{\sin \sqrt{\pi^\dagger \pi}}{\sqrt{\pi^\dagger \pi}} \end{bmatrix} \quad (8.48)$$

Exercise *Coordinates and coordinate patches*

Write out an explicit system of coordinates and coordinate patches for $Gr_{2,n}$.

Exercise *Projection operators*

An important remark for later chapters is that we can view $Gr_{k,n}$ as the space of orthonormal projection operators on $\mathbb{F}^n$ of rank k .

- a.) Explain why this is so.
- b.) Explain the symmetry under $k \rightarrow (n - k)$ in the above formulae.
- c.) Show that, if Λ is a matrix made from an ON basis of V then we can identify

$$P = \Lambda^\dagger \Lambda \quad (8.49)$$

d.) Show that, for $Gr_{1,2}(\mathbb{C})$ the patch around V_0 is described by the family of projection operators

$$P(z, \bar{z}) = \frac{1}{1 + |z|^2} \begin{pmatrix} 1 & z \\ \bar{z} & |z|^2 \end{pmatrix} \quad (8.50)$$

- e.) Show that, if $P(t)$ is a differentiable family of projection operators then

$$\dot{P}P = (1 - P)\dot{P} \quad (8.51)$$

and use this to interpret $\dot{P}$ as a map from V to $V^\perp$.

- f.) Show that $T_V Gr_{k,n} \cong \text{Hom}(V, V^\perp)$.
- g.) Show that a natural metric on the Grassmannian is

$$(A_1, A_2) = \text{Tr}(A_1^\dagger A_2) \quad (8.52)$$

and that this agrees with the homogeneous metric obtained using the Cartan-Killing form on $SU(n)$. Express the metric in terms of projection operators.

Exercise *Flag varieties*

Consider the space of flags:

$$\mathcal{F}_{123} = \left\{ V_1 \subset V_2 \subset V_3 = \mathbb{C}^{k_1+k_2+k_3=N} \right\} \quad (8.53)$$

where

$$\begin{aligned} \dim V_1 &= k_1 \\ \dim V_2 &= k_1 + k_2 \end{aligned} \quad (8.54)$$

- a) Show that this space is the homogeneous space

$$\frac{U(k_1 + k_2 + k_3)}{U(k_1) \times U(k_2) \times U(k_3)} \quad (8.55)$$

b) Show that it is a fiber bundle over $Gr_{k_1+k_2}(\mathbb{C}^{k_1+k_2+k_3})$. Find the fiber.

This makes precise the idea of constructing a flag by choosing a subspace V_2 of $\dim = k_1 + k_2$ and then a subspace $V_1 \subset V_2$ of $\dim k_1$.

Is this space a product of Grassmannians? (This is a hard question.)

c.) The full flag variety in $\mathbb{C}^N$ is the space of all flags $V_1 \subset V_2 \subset \cdots \subset V_N$ of increasing dimension where $\dim V_i = i$. Show that this is G/T where $G = U(N)$ and T is a maximal torus $T \cong U(1)^N$.

8.1.4 Schubert cells

Recall that every $k \times n$ matrix can be put into *reduced row echelon form*, by Gaussian elimination, which consists of elementary row operations.⁶³ These operations generate all of $GL(k)$ and hence, if $[\Lambda]$ is in U_J we know it can be put into reduced row echelon form.

There are different conventions for reduced row echelon form. We will take it to be a lower triangular matrix, with the leading coefficient of one row strictly to the left of the next lower row. The reduced form requires that the column vector associated with the leading entry in any row have all entries zero except for that row, in which case the entry is $= 1$. The columns corresponding to the leading nonzero entry of any row determine a Schubert symbol.

Example 1: The matrix

$$\begin{pmatrix} \mathbf{1}_{k \times k} & 0_{k \times (n-k)} \end{pmatrix} \quad (8.56)$$

is in reduced row echelon form.

Example 2 In $Gr_{3,8}$, if $J = \{2, 4, 7\}$ then $\Lambda \in U_J$ can be put in the form

$$\Lambda = \begin{bmatrix} * & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & 0 & * & 1 & 0 & 0 & 0 & 0 \\ * & 0 & * & 0 & * & * & 1 & 0 \end{bmatrix} \quad (8.57)$$

Example 2 In $Gr_{5,10}$, if $J = \{1, 3, 6, 9, 10\}$ then $\Lambda \in U_J$ can be put in the form

$$\Lambda = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & * & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & * & 0 & * & * & 1 & 0 & 0 & 0 & 0 \\ 0 & * & 0 & * & * & 0 & * & * & 1 & 0 \\ 0 & * & 0 & * & * & 0 & * & * & 0 & 1 \end{bmatrix} \quad (8.58)$$

⁶³These are exchange of rows, multiplication of a row by a nonzero scalar, and the addition of one row to another row.

The set of matrices in reduced row echelon form for J will be called X_J^0 . Notice that it has dimension

$$\dim X_J = \sum_{i=1}^k (j_i - i) \quad (8.59)$$

For example in (8.57) we count the number of undetermined coefficients *. For $i = 1$ we have $j_1 = 2$ and just one coefficient. In general if the last nonzero answer is in column j_i there would be j_i coefficients in that row. However, in reduced row echelon form the last nonzero entry is fixed to be one and the previous $(i - 1)$ columns for j_s with $s < i$ are fixed to have a zero. Hence there are only $\lambda_i = j_i - i$ free parameters.

Note that for the case $j_i = i$ we get a single point. While the largest space occurs for

$$j_1 = n - k + 1 < j_2 = n - k + 2 < \cdots < j_k = n - k + k = n \quad (8.60)$$

and has dimension $k(n - k)$.

The sets X_J^0 have closures X_J known as *Schubert cells*. There are many beautiful facts about these cells:

1. They provide a cell decomposition of the Grassmannian:

$$Gr_{k,n} = \coprod_J X_J^0 \quad (8.61)$$

2. The cells are enumerated by Young diagrams: Indeed, note that $\lambda_i = j_i - i$ is a nondecreasing sequence of positive integers, and is hence a partition. In fact, the pattern of stars in (8.57) outlines a Young diagram!
3. The homology classes $[X_J]$ forms a basis for the homology group of the Grassmannian. Moreover, the intersection theory in this basis can be expressed very elegantly.

Example 1 If we consider $\mathbb{CP}^1 = Gr_{1,2}(\mathbb{C})$ then there are two Schubert cells. The Schubert symbol $J_2 = \{2\}$ corresponds to $\Lambda = \begin{pmatrix} * & 1 \end{pmatrix}$. Now $X_{J_2}^0$ is a copy of $\mathbb{C}$, corresponding to all of the Riemann sphere except a point at infinity. The other Schubert symbol is $J_1 = \{1\}$ and corresponds to $\Lambda = \begin{pmatrix} 1 & 0 \end{pmatrix}$. This is the point at infinity.

Example 2 This generalizes to $\mathbb{CP}^n = Gr_{1,n+1}(\mathbb{C})$. Now for $J_{n+1} = \{n+1\}$ we have $\Lambda = \begin{pmatrix} * & \cdots & * & 1 \end{pmatrix}$ and the cell is just $\mathbb{C}^n$. It misses the $\mathbb{CP}^{n-1}$ at infinity. Altogether the decomposition has one cell in each complex dimension n .

Example 3 The first new example beyond projective spaces is $Gr_{2,4}$...

EXPLAIN DETAILS

9. Bundle Basics

Definition A G -action on X is *simply transitive* if it is both transitive and free. In this case X is said to be a *principal homogeneous space* or a G -torsor.

To make clear the distinction between a group G and a G -torsor consider the set of real points $\mathbb{Z} + \theta$ where θ is some real number, not necessarily integral. This is clearly a $\mathbb{Z}$ -torsor, but is not a group when θ is not an integer, since there is no natural zero.

If G is a topological group then continuous families of G -torsors are called *principal bundles*. A primary example is a product space $B \times G$.

To give a formal definition we first need the general definition of a *fiber bundle*:

Definition: Let E, B be topological spaces, $p : E \rightarrow B$ be a continuous projection of topological spaces. Then p is said to be *locally trivial* if there is a topological space F such that, for every point $b \in B$ there is some neighborhood $b \in U \subset B$ so that $p^{-1}(U)$ is homeomorphic to $U \times F$. That is

$$\begin{array}{ccc} p^{-1}(U) & \xrightarrow{\varphi} & U \times F \\ \downarrow p & \nearrow p_1 & \\ U & & \end{array} \quad (9.1)$$

where p_1 is the projection on the first factor.

The space F is called the *fiber space* of the fiber bundle. The preimage $p^{-1}(b)$ of any point $b \in B$ is called the *fiber over b* and is sometimes denoted E_b . All the spaces E_b are homeomorphic to F .

STRUCTURE GROUP OF A BUNDLE

DEFINITION OF EQUIVALENCE OF BUNDLES, TRIVIAL BUNDLE, TRIVIALIZABLE BUNDLE

Definition If a fiber bundle $p : P \rightarrow B$ has a fiber space which is a topological group G and there is a free right G -action on P then it is said to be a *principal G -bundle over B* .

The fibers $p^{-1}(b)$ of a principal fiber bundle are examples of principal homogeneous spaces for G - there is no natural choice of the identity element. In fact, the principal bundle is equivalent (in a suitable sense) to the trivial bundle $P = B \times G$ iff there is a continuous cross-section $s : B \rightarrow P$.

Figure 59: The projection $\pi : \mathbb{R} \rightarrow S^1$ defines a nontrivial principal $\mathbb{Z}$ -bundle over the circle.

Example 1. Consider the projection $\pi : \mathbb{R} \rightarrow S^1$ given by $\pi(x) = e^{2\pi i x}$. For each point on the circle there is a $\mathbb{Z}$ -torsor $\mathbb{Z} + x$ which is a principal homogeneous space. Note however, that the disconnected fibers fit together to make a nice connected “total space” $\mathbb{R}$. The picture one should have is of an infinite spiral covering the circle as in 59. The fiber above $e^{2\pi i \theta}$ is the $\mathbb{Z}$ torsor we defined above $\mathbb{Z} + \theta$. Note that if one insisted on picking a zero at some particular θ_0 , then smoothly continuing this zero as $\theta_0 \rightarrow \theta_0 + 2\pi$ we would get a different zero. This illustrates cleanly that there is no natural choice of zero.

Example 2.

$$\pi : SO(3) \rightarrow SO(3)/SO(2) \cong S^2. \quad (9.2)$$

Example 3.

$$\pi : GL(n, \mathbb{C}) \rightarrow GL(n, \mathbb{C}) / (GL(k, \mathbb{C}) \times GL(n - k, \mathbb{C})) \cong Gr(k, n). \quad (9.3)$$

Exercise

Show that a principal bundle is trivializable iff it has a continuous global cross section.

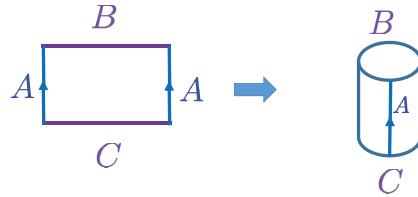


Figure 60: Identifying opposite sides of a rectangle with indicated orientations gives a cylinder. Draw open sets near the vertical line on the cylinder and their inverse images in the rectangle.

10. The classification of compact two-dimensional surfaces

A basic and beautiful theorem is the classification of compact two-dimensional surfaces. It is used all the time in string theory, as well as in other areas of physical mathematics.

By a *surface* we mean a two-dimensional topological manifold, possibly with boundary. See Section 6 below for the official definition of a manifold. In this section we will be very informal: A surface is a topological space that locally “looks like” $\mathbb{R}^2$ or the upper half-plane. The classification of surfaces up to homeomorphism is a beautiful chapter

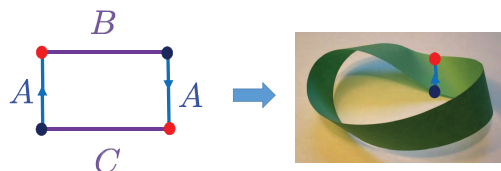


Figure 61: Identifying opposite sides of a rectangle with indicated orientations gives a Möbius strip. Figure of Möbius strip from Wikipedia.

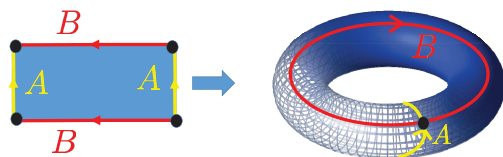


Figure 62: Identifying both opposing sides of a rectangle, with the indicated orientations gives a torus.

in topology. It goes back to the 1860's and has a colorful history. A fully accepted mathematical proof was not available until the first half of the twentieth century. Good references for the classification proof are:

1. W.S. Massey, *Algebraic Topology: An Introduction*, Springer GTM 56
2. Seifert and Threlfall
3. M.A. Armstrong, *Basic Topology*, McGraw Hill, 1979.

Here are some nontrivial statements that a careful treatment would prove in detail: It turns out that all surfaces can be cut up into polygons, and hence into triangles, and then glued back together. If the boundaries of the triangles can all be coherently oriented then the surface is said to be *orientable*. If some pair of triangles is always frustrated then it is said to be *unorientable*. This notion does not depend on the choice of triangulation. Moreover, the *Euler character* of the surface may be defined to be $\chi = V - E + F$, where V is the number of vertices, E the number of edges, and F the number of faces. This number does not depend on the triangulation. See Section 14.1 below for more discussion of the Euler character in more generality.

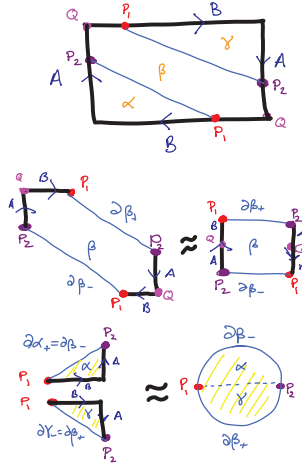


Figure 63: This figure indicates why sewing together a Mobius strip and a disk along the circular boundaries is a surface homeomorphic to $\mathbb{R}P^2$. Identifying the sides of the rectangle above according to the indicated orientations gives a copy of $\mathbb{R}P^2$. (See Figure *** below.) On the other hand we can cut the rectangle into three pieces, denoted α, β, γ . The boundary identifications on α give a Mobius strip whose bounding circle is the path $\partial\beta_+ \cup \partial\beta_-$ from p_2 to p_3 and back to p_2 . The pieces α, γ can be reassembled by first gluing along the pieces of the A and B side. The result is a disk the boundary of this disk is to be identified with the boundary of the Mobius strip from α .

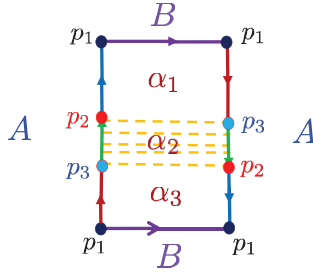


Figure 64: The identifications in this figure define the Klein bottle. If we cut the rectangle into three subrectangles $\alpha_1, \alpha_2, \alpha_3$, where α_2 is the golden region, then α_2 is a Mobius strip with boundary circle $p_2p_3p_2$. Gluing the rectangles α_1 to α_3 along side B produces another Mobius strip with boundary circle $p_2p_3p_2$. Therefore, the Klein bottle can be presented as two Mobius strips sewn along their common circle.

Let us start the discussion by considering a disk, which is homeomorphic to a rectangle. We can identify sides of a rectangle in various interesting ways. First, as in 60 we can produce a cylinder, or, equivalently, a sphere with two disks cut out. If we glue together the two ends of a cylinder we get a torus. Therefore, identifying the sides of the rectangle as in Figure 62 produces a torus. When we produce a torus this way there are two distinguished

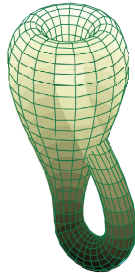


Figure 65: A Klein bottle immersed in $\mathbb{R}^3$. Figure from Wikipedia.

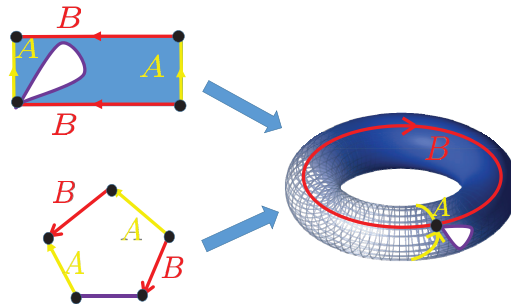


Figure 66: Cutting a disk out of a torus.

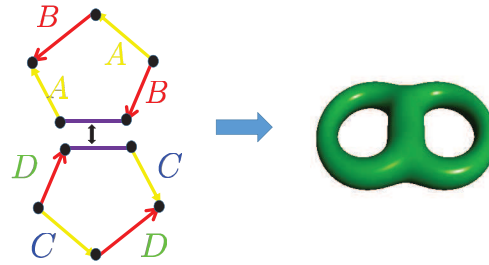


Figure 67: Gluing together copies of a torus with one hole gives a genus two surface.

closed paths called the A-cycle and the B-cycle.

Returning to the construction of the cylinder, we can change the arrows in the identification to produce a Mobius strip as in Figure 61. Now, the boundary of a Mobius strip is a circle. The boundary of a disk is also a circle. What happens when we identify these two

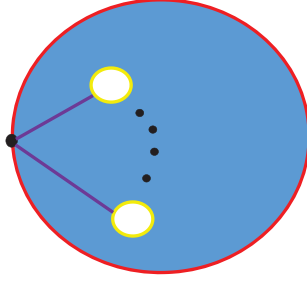


Figure 68: Identify the red boundary of the disk to a point and cut along the purple lines (or identify them to the same point) to get a sphere with holes cut out. The number of holes is the number of yellow circles.

circles? We claim the resulting surface is homeomorphic to the projective plane. To see this, note that the identifications of the boundary of a disk producing the projective plane, such as shown in Figure 69 can be deformed to the identifications of the boundary of a rectangle shown in Figure 63. Then, as explained in that figure, by cutting up the diagram and regluing we can see that it is the same (“same” means “same homeomorphism type”) as a disk sewn to a Mobius strip.

The final surface we can get from the rectangle is the Klein bottle. This is shown in Figure 64. By cutting and pasting as indicated in that figure we can identify the Klein bottle as homeomorphic to a sphere with two disks cut out and two Mobius strips sewn in. On the other hand if we first sew together sides B to get a cylinder and then identify the two ends of the cylinder with the orientations dictated by the diagram we can almost picture this surface immersed in $\mathbb{R}^3$, as shown in the famous figure 65.

Recall from Figure 62 that a torus can be made by identifying opposite sides of a rectangle. Now, cut a disk out of the torus to produce a handle: The handle can be viewed as a disk with two disks cut out, and the ends of the cylinder glued back in as in the top figure in 71. (This latter representation makes clear that the name is apt.) Viewing the torus as a rectangle with identifications we thus represent a handle as a pentagon with identifications as in Figure 66.

Glue together two handles along their common circle. Using the pentagon representation we get an octagon with sides identified by

$$A_1 B_1 A_1^{-1} B_1^{-1} A_2 B_2 A_2^{-1} B_2^{-1} \quad (10.1)$$

See Figure 67

A *genus g surface* is a surface obtained by cutting g disks out of a sphere and gluing a handle back into each disk. We can find a nice presentation by representing the two-dimensional sphere as a disk with the boundary identified to a point. Then the sphere with g holes can be represented as in 68. We are free to shrink the cuts down to a point. When we glue in a handle into each hole we get a presentation of the genus g surface as a

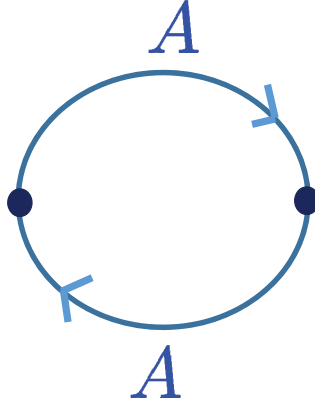


Figure 69: Identification of the boundary of a disk to produce the projective plane can be thought of as the identification of the boundary of a 2-gon with symbol AA .

$4g$ -gon with identifications of the sides according to a sequence of 4 adjacent sides identified according to $A_j B_j A_j^{-1} B_j^{-1}$ where the inverse means the direction of the arrow is reversed. Thus we have the gluing of the boundaries of a $4g$ -gon encoded in the word:

$$A_1 B_1 A_1^{-1} B_1^{-1} \cdots A_g B_g A_g^{-1} B_g^{-1} \quad (10.2)$$

All of these surfaces are orientable.

Similarly, we can take a sphere with n holes and glue in a Mobius strip along its S^1 boundary on each of the holes. Now, if we view the projective plane as the closed disk with antipodal points on the boundary identified then we can view it as the polygon with 2 sides glued as in Figure 69. Cutting out a disk, as with a handle, and gluing into 68 where there are n holes we obtain a $2n$ -gon encoded in the word:

$$A_1 A_1 A_2 A_2 \cdots A_n A_n \quad (10.3)$$

None of these surfaces are orientable, if $n > 0$.

We have now produced many surfaces. It is not trivial, but true that all compact connected surfaces with boundary can be obtained by cutting out disks from a sphere and sewing in handles, or Mobius strips, or nothing.

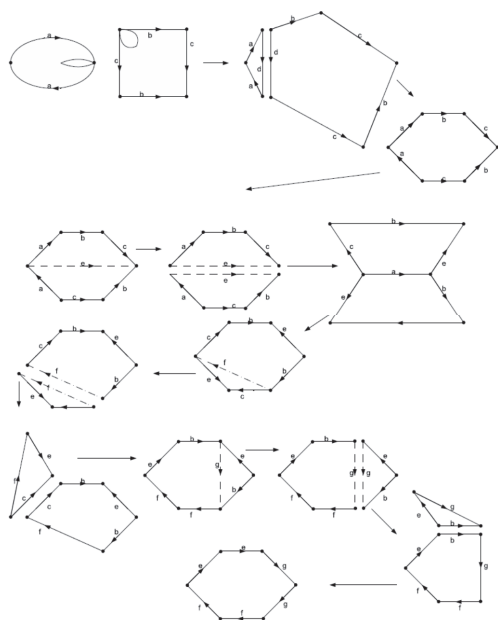


Figure 70: By cutting and reassembling we can convert a connected sum of a projective plane and a torus into a connected sum of three projective planes.

What happens if we glue in *both* Mobius strips and handles? We now claim that if we consider a sphere with three holes cut out and 3 Mobius strips glued in we get the same surface as a sphere with one handle and one Mobius strip glued in. Thus, as long as there is one Mobius strip glued in we can convert handles to pairs of Mobius strips.

One proof, taken from ⁶⁴ is obtained by a clever cut and paste manipulation of the surfaces presented as identified polygons. It is illustrated in Figure 70. While clever, it is not extremely intuitive. Another argument, adapted from Armstrong's book, pp. 149-152 is summarized in Figures 71 and 72.

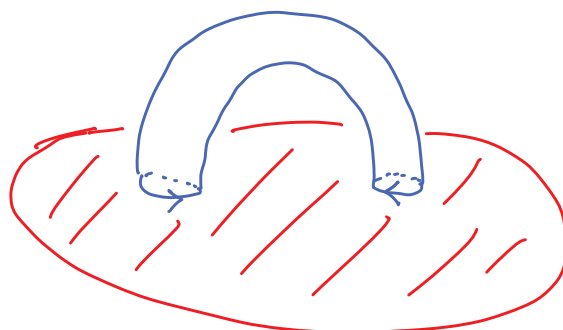
Putting these facts together, (together with some further reasoning which one can find, for example, in. ⁶⁵) we get the beautiful classification theorem:

Theorem: Any compact connected two-dimensional surface (possibly with boundary) is classified up to homeomorphism by three pieces of data:

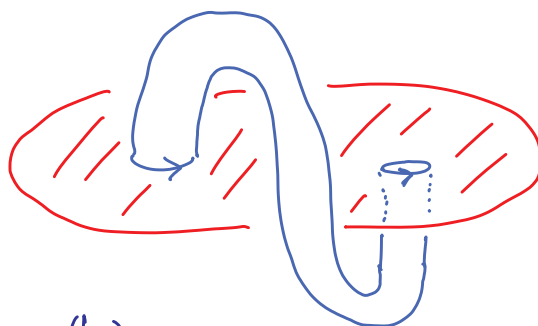
1. Whether it is orientable or not.

⁶⁴J. Huang, "Classification of Surfaces," <http://www.math.uchicago.edu/~may/VIGRE/VIGRE2008/REUPapers/Huang.pdf>

⁶⁵W. Massey, *Introduction to Algebraic Topology*.



(a): Torus - Disk



(b): Klein Bottle - Disk



(c): Mobius strip

Figure 71: (a) This surface is obtained by cutting out a disk from a sphere with a handle attached. (b): This surface is obtained by cutting out a disk from a Klein bottle. Think of the disk that has been cut out as the one which is responsible for the self-intersections in the immersion into $\mathbb{R}^3$. The difference between (a) and (b) is the relative orientations of the attached ends of the cylinder. (c) Now, we want to sew in a Mobius strip replacing the disk which was removed from the sphere in (a) and (b) (and is not shown in (a) and (b)).

2. The number of boundary components
3. Its Euler character. (See below.)

Equivalently, we can say that any compact connected two-dimensional surface (possibly with boundary) is homeomorphic to

1. A two-dimensional sphere with b disks cut out and g handles attached.

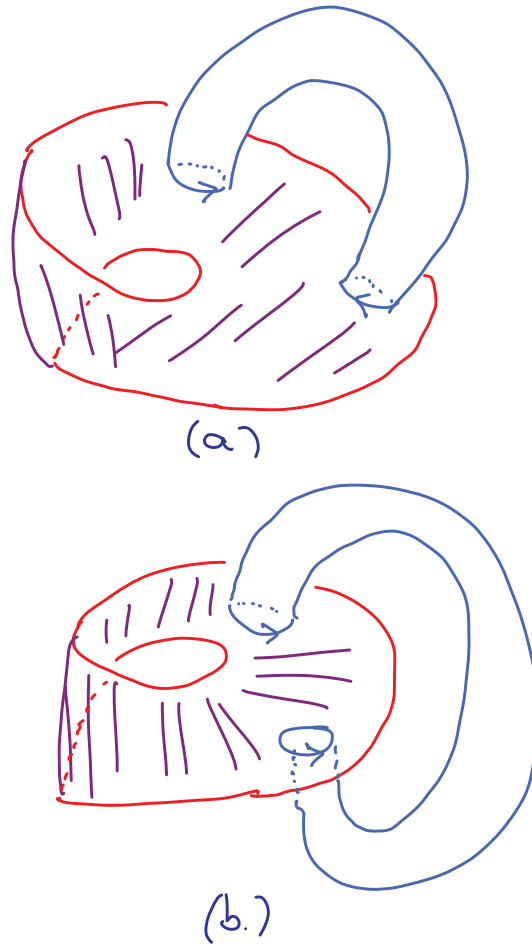


Figure 72: After gluing in the the boundary of the Möbius strip (c) in the previous figure to the boundary of (a) or (b) we can move the ends of the cylinder glued to (a) or (b) onto the Möbius strip to produce figures (a) and (b) above. These are to be sewed into an ordinary disk. On the other hand, we can now turn (a) into (b) by sliding one end of the cylinder around the the Möbius strip!

2. A two-dimensional sphere with b disks cut out and n Möbius strips attached, where $n > 0$.

Moreover, no two of the above surfaces are homeomorphic.

Remarks

1. A projective plane with a disk removed is called a *crosscap*. A surface is represented as a sphere with b boundaries, with g then handles attached, and c crosscaps attached then the Euler character is:

$$\chi = 2 - 2g - b - c \quad (10.4)$$

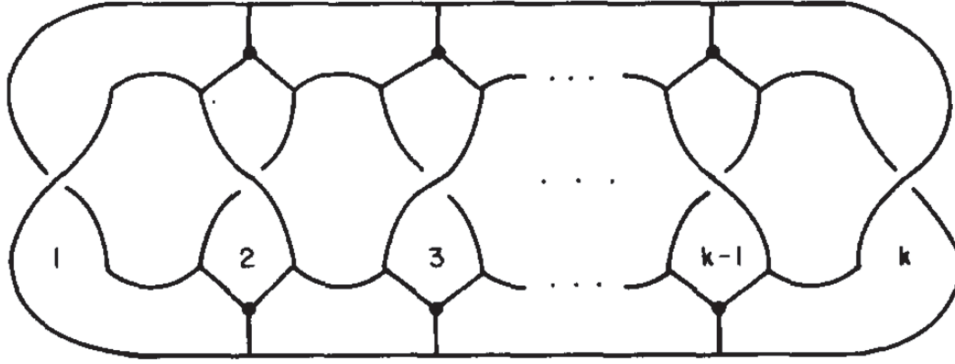


Figure 73: A surface. Note that the vertical bands are twisted in alternating ways as we read horizontally.

Exercise

Consider Figure 73. ⁶⁶ Find the homeomorphism type of this surface in the classification theorem. ⁶⁷

b.) Show that the two surfaces in Figure 74 are homeomorphic. ⁶⁸

⁶⁶The figure is taken from Fig. 3 of S.B. Giddings, E. Martinec, and E. Witten, “Modular Invariance in String Field Theory,” *Physics Letters* **176B** (1986) 362.

⁶⁷*Answer:* The surface is orientable since any closed loop flips the orientation twice. For k odd (which is the case illustrated) there is a single boundary. The analogous surface for k even has two boundaries. Now compute $V = 8(k - 2)$, $E = 12(k - 2)$, and $F = 3(k - 2)$. Now $V - E + F = 2 - k = 2 - 2g - b$, so $2g + b = k$. Therefore, for k odd, $g = (k - 1)/2$ and for k even $g = (k - 2)/2$.

⁶⁸The figure is taken from P. Aspinwall, T. Bridgeland, et. al. *Dirichlet Branes and Mirror Symmetry*, pp. 37-38.

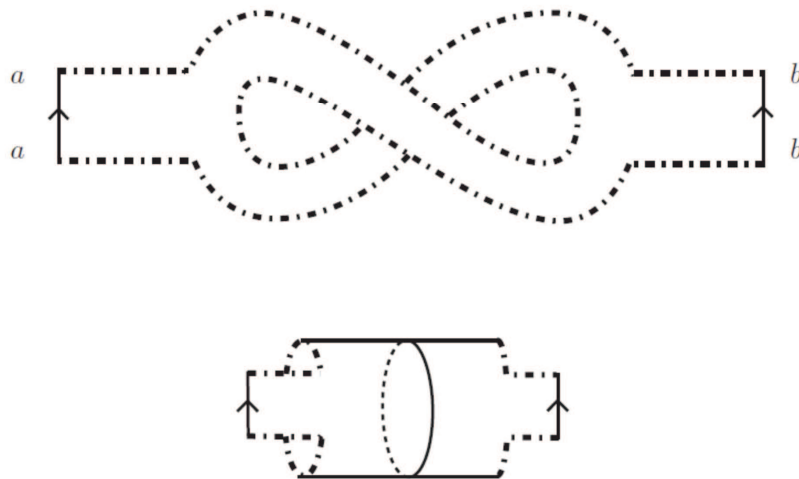


Figure 74: Exercise: Show that the surfaces shown here are homeomorphic.

Remark: The surface shown in (a) has useful applications to the cellular decomposition of the moduli space of Riemann surfaces and to string field theory. The equality of the surfaces in (b) has important implications in topological field theory and in string theory since it relates a “purely open string process” to an “exchange of a closed string.” The top figure in Figure 74 represents a one-loop diagram in open string field theory. (If we shrink the width of the strips to zero we get a standard one-loop diagram with two external lines and two trivalent vertices.) On the other hand, the lower figure represents a *tree level* diagram in which an open string closes on itself, propagates as a closed string, and then splits into another open string. The fact that the classical tree-level diagram of the open-closed theory is the same as the quantum one-loop diagram in the open string theory has some deep and surprising implications in string theory.

Exercise *Gluing*

Consider a torus obtained by gluing together opposite sides of a square. Define closed oriented curves A and B as in the notes. They are usually called *A-cycles* and *B-cycles*.
a.) *S-transformation*. Centering the square on the origin, consider the transformation of

the torus to itself by making a 90 degree counterclockwise rotation before gluing. Describe what this does to the standard A and B cycles. Describe the action of S^2 .

b.) *Dehn twist.* Let $\gamma \subset C$ be a closed curve in a surface. A Dehn twist around γ is a diffeomorphism of C to itself obtained by isolating a small annular neighborhood of γ and twisting one circle boundary of the cylinder by 2π , holding the other circle boundary fixed. Draw pictures illustrating the action of a Dehn twist around the A -cycle on both the A -cycle and the B -cycle.

c.) Consider the operation of first doing a Dehn twist around the A -cycle and then doing an S -transformation. Call this ST . What happens if we perform ST three times?

d.) Consider two solid tori, considered as a product of a disk and a circle $D^2 \times S^1$. What three-dimensional manifold is obtained by gluing together the solid tori along their torus boundary where one first makes an S -transformation on one of the two tori?

Exercise Klein bottle from group actions

a.) Show that the Klein bottle is a free quotient of the torus. Let (σ^1, σ^2) with $\sigma^i \sim \sigma^i + 1$ be coordinates on the torus T^2 . Consider the fixed-point free involution $(\sigma^1, \sigma^2) \rightarrow (\sigma^1 + \frac{1}{2}, 1 - \sigma^2)$.

b.) Show that the Klein bottle is a quotient of $\mathbb{R}^2$ by the group generated by the operations

$$\begin{aligned} g_1 : (\sigma_1, \sigma_2) &\rightarrow (L - \sigma_1, \sigma_2 + \beta) \\ g_2 : (\sigma_1, \sigma_2) &\rightarrow (\sigma_1 + L, \sigma_2) \end{aligned} \tag{10.5}$$

Show that these satisfy the relation $g_2 g_1 g_2 = g_1$.

11. Homotopy of maps and spaces

11.1 Homotopy of maps

Recall that in Section **** we defined a homotopy of two continuous maps $f_0, f_1 \in \mathcal{C}(X, Y)$ as a continuous path $\varphi : [0, 1] \rightarrow \mathcal{C}(X, Y)$ (in the compact-open topology for $\mathcal{C}(X, Y)$) that connects them. An equivalent and perhaps more standard definition is:

Definition. Two continuous maps of topological spaces:

$$\begin{aligned} f_0 : X &\rightarrow Y \\ f_1 : X &\rightarrow Y \end{aligned} \tag{11.1}$$

are *homotopic* if there is a continuous map $F : [0, 1] \times X \rightarrow Y$ such that

$$F(0, x) = f_0(x) \quad \& \quad F(1, x) = f_1(x) \tag{11.2}$$

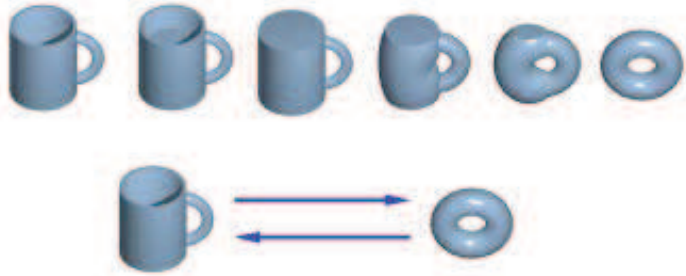


Figure 75: The surface of a donut or of a mug can be viewed as a map from $S^1 \times S^1$ into $\mathbb{R}^3$. These two maps are related by a homotopy. This figure was taken from http://inperc.com/wiki/index.php?title=Homotopy_and_homotopy_equivalence.

We say that F is a *homotopy* between f_0 and f_1 . If we think of a continuous path in $\mathcal{C}(X, Y)$ then $\varphi(t)$ is the function $f_t : X \rightarrow Y$ given by $F(t, \cdot)$.

Example: For an example see Figure 75.

Definition

a.) A continuous map $f : X \rightarrow Y$ is a *homotopy equivalence* of spaces X and Y if there is a continuous map $g : Y \rightarrow X$ such that $f \circ g : Y \rightarrow Y$ is homotopic to Id_Y and $g \circ f : X \rightarrow X$ is homotopic to Id_X . Such a map g is called a *homotopy inverse* of f .

An important special case of homotopy equivalence is given by the concept of a *deformation retract*:

Definition: Let A be a subspace of a topological space X . Let $\iota : A \hookrightarrow X$ be the inclusion map: $\iota(a) = a$, for all $a \in A$.

a.) A continuous map $r : X \rightarrow A$ such that $r \circ \iota = Id_A$ is called a *retract*. That is, $r(a) = a$ for all $a \in A$.

b.) If $\iota \circ r : X \rightarrow X$ is homotopy equivalent to Id_X then r is called a *deformation retract*.

c.) If there is a deformation retract of a space X to a point in X it is said to be *contractible*.

Examples:

1. $\mathbb{R}^n$ is homotopy equivalent to a point, so it is contractible. This shows that homotopy equivalence of spaces is weaker than homeomorphism. Exercise: Prove this in detail.
2. There is a deformation retract of $\mathbb{R}^{n+1} - \{\vec{0}\}$ to S^n . Again: Prove this in detail.
3. $S^n, \mathbb{RP}^n$ are not contractible spaces. How would you prove this?
4. *Path space*: Let X be any path-connected space and $x_0 \in X$. Let $\mathcal{P}(X, x_0)$ be the space of all continuous paths $\varphi : [0, 1] \rightarrow X$ with $\varphi(0) = x_0$. Topologize $\mathcal{P}(X, x_0)$ with the compact-open topology. This topological space is known as the *path space* of X (based at x_0). We claim this is a contractible space. What is the contracting homotopy? ⁶⁹

Exercise

- a.) Show that homotopy of maps is an equivalence relation.
- b.) Show that the equivalence classes are the path-connected components of $\mathcal{C}(X, Y)$ in the compact-open topology. Recall that $\pi_0(\mathcal{C}(X, Y))$ is also denoted $[X, Y]$.
- c.) Show that if X or Y is contractible then $[X, Y]$ consists of one element.
- d.) Show that A is a deformation retract of X if there exists a continuous map $R : [0, 1] \times X \rightarrow X$ such that

1. $\forall x \in X, H(0, x) = x.$
2. $\forall x \in X, H(1, x) \in A.$
3. $\forall a \in A, t \in [0, 1], H(t, a) = a$

- e.) Write the function R for the deformation retract of $\mathbb{R}^n$ to the origin and of $\mathbb{R}^{n+1} - \{\vec{0}\}$ to S^n .
-

⁶⁹ Answer: $F(\varphi, s)$ is the path defined by $\varphi_s(t) := \varphi(st)$. That is φ_s is the path that runs along $\varphi(t)$ just up to $\varphi(s = t)$.

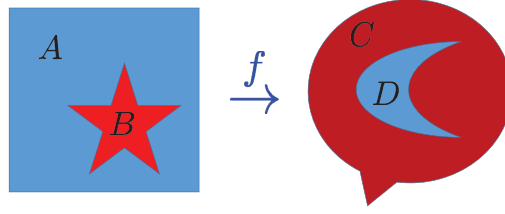


Figure 76: A map of pairs. The function f takes the blue square A into the maroon bubble C and takes the subspace B given by the red star into the blue moon D .

11.2 Homotopy of maps of pairs

It is often useful to refine the notion of homotopy to *relative homotopy*. To define this we first define a *map of pairs*, denoted by:

$$(A, B) \xrightarrow{f} (C, D) \quad (11.3)$$

to be a map $f : A \rightarrow C$ which also takes a subset $B \subset A$ into a subset $D \subset C$, that is, $f(B) \subset D$. See Figure 76.

Two maps $f_0, f_1 : (A, B) \rightarrow (C, D)$ are called *relatively homotopic* if the homotopy $F(t, x)$ between them also takes $B \rightarrow D$ and is *constant* in t :

$$F(t, x) = f_0(x) = f_1(x) \quad x \in B \quad (11.4)$$

Again relative homotopy defines an equivalence relation and we will denote the set of homotopy classes by $[(A, B), (C, D)]$. There is an analogous formula to (4.11).

Definition A space X with a distinguished point, called the *basepoint*, is called a *pointed space*. The set of homotopy classes of maps between two pointed spaces

$$[(X, x_0), (Y, y_0)] \quad (11.5)$$

are sometimes just written as $[X, Y]_*$

Exercise

a.) Show that if $g : Y \rightarrow Z$ takes $y_0 \in Y$ to $z_0 \in Z$ then there is a map

$$g_* : [(X, x_0), (Y, y_0)] \rightarrow [(X, x_0), (Z, z_0)] \quad (11.6)$$

defined by

$$g_*([f]) := [g \circ f] \quad (11.7)$$

b.) Show that if g is a homotopy equivalence then g_* is a bijection.

11.2.1 Example: Homotopy of curves

Let us unpack the above definition for the very important special case of homotopy of two curves relative to endpoints. If $p_0, p_1 \in X$ are two points then let us write out in detail the meaning of the homotopy used to define the equivalence classes in

$$[(0, 1], \{0, 1\}), (X, \{p_0, p_1\})] \quad (11.8)$$

Thus, we consider two continuous curves γ_0, γ_1 with the same endpoints:

$$\begin{aligned} \gamma_0(0) &= \gamma_1(0) = p_0 \\ \gamma_0(1) &= \gamma_1(1) = p_1 \end{aligned} \quad (11.9)$$

These are homotopic with fixed endpoints in a topological space X , written

$$\gamma_0 \sim \gamma_1 \quad (11.10)$$

if there is a continuous function

$$\Gamma : [0, 1] \times [0, 1] \rightarrow X \quad (11.11)$$

such that

$$\begin{aligned} \Gamma(t, 0) &= \gamma_0(t) \\ \Gamma(t, 1) &= \gamma_1(t) \\ \Gamma(0, s) &= p_0 \\ \Gamma(1, s) &= p_1 \end{aligned} \quad (11.12)$$

This is illustrated in 77.

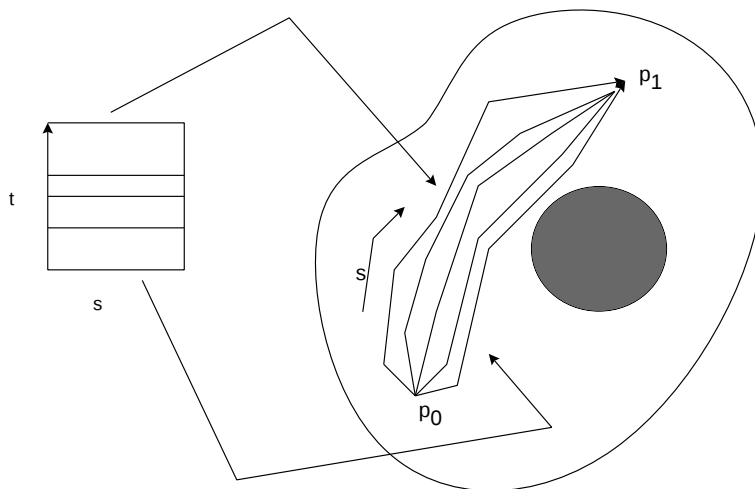


Figure 77: Think of t as a continuous time-development of the curves defined by $s \in [0, 1]$. REDO
FIGURE: s and t ARE BACKWARDS FROM THE TEXT!

Exercise

Show that $\gamma_0 \sim \gamma_1$ is an equivalence relation.

12. Homotopy groups

In general the set $[(A, B), (C, D)]$ is just a set with no extra natural structure. However, in some cases where (A, B) is a pointed sphere and (C, D) is a pointed space these sets can be given the structure of groups, and these groups are very important “measures” of the topological complexity of the space C .

12.1 π_0

$\pi_0(X)$ is the set of connected components of X . In general it has no natural group structure.

However, if G is a topological group then the connected component of the identity, G_0 , is a normal subgroup. Then $\pi_0(G) = G/G_0$. Assuming the group is locally path-connected it is easy to see that G_0 is a normal subgroup, for if $g \in G_0$ then there is a continuous path $\varphi : [0, 1] \rightarrow G_0$ with $\varphi(0) = 1$ and $\varphi(1) = g$. Then for any $h \in G$, the path $t \mapsto h\varphi(t)h^{-1}$ must also be in G_0 , because there is a homotopy $F(t, s) = h\varphi(st)h^{-1}$ with $s \in [0, 1]$ to the constant path. Therefore, $hgh^{-1} \in G_0$, and G_0 is normal. It can be shown that G/G_0 is always a discrete group.

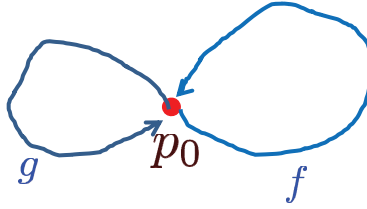


Figure 78: Two loops f, g with basepoint at p_0 .

12.2 The fundamental group: π_1

Choose a point $x_0 \in X$. The fundamental group $\pi_1(X, x_0)$ based at x_0 is, as a set, the set of homotopy classes of closed curves. That is we consider maps:

$$f : ([0, 1], \{0, 1\}) \rightarrow (X, \{x_0\}) \quad (12.1)$$

and, as a set, we define $\pi_1(X, x_0)$ to be the set of homotopy equivalence classes of such maps of pairs. As usual they are denoted $[f]$.

We can define a group structure on $\pi_1(X, p_0)$ by concatenating curves as in 79 and rescaling the time variable so that it runs from 0 to 1. In equations we have

$$f_1 \star f_2(t) := \begin{cases} f_1(2t) & 0 \leq t \leq \frac{1}{2} \\ f_2(2t - 1) & \frac{1}{2} \leq t \leq 1 \end{cases} \quad (12.2)$$

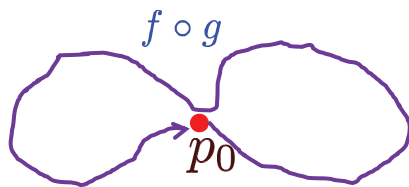


Figure 79: The concatenation of the loops $f \star g$. Note that the “later” loop is written on the right. This is generally a more convenient convention when working with homotopy and monodromy. In order for $f \star g$ to be a map from $[0, 1]$ into X we should run each of the individual loops at “twice the speed” so that at time $t = 1/2$ the loop returns to p_0 . However, in homotoping $f \star g$ there is no reason why the point at $t = 1/2$ has to stay at p_0 .

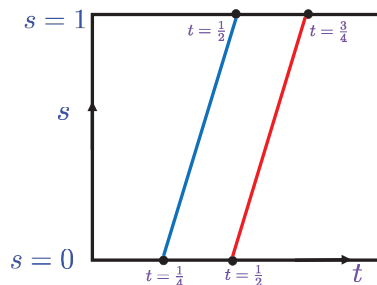


Figure 80: The homotopy demonstrating that loop concatenation is an associative multiplication on homotopy equivalence classes of closed loops. The blue line is $s = 4t - 1$ and the red line is $s = 4t - 2$.

Remarks

1. Note that we are composing successive paths *on the right*. This is slightly nonstandard but a nice convention when working with monodromy and path ordered exponentials of gauge fields - one of the main physical applications.
2. Note well that $(f_1 \star f_2) \star f_3$ is *NOT* the same path as $f_1 \star (f_2 \star f_3)$. We will comment much more on that in Section 12.2.4
3. For the moment we simply notice that if we mod out by homotopy then we have a well-defined product on homotopy classes in $\pi_1(X, x_0)$

$$[f_1] \cdot [f_2] := [f_1 \star f_2] \tag{12.3}$$

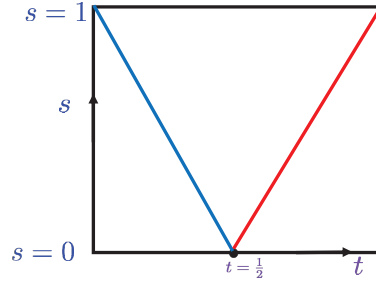


Figure 81: The homotopy demonstrating that the loop $g(t) = f(1 - t)$ provides a representative for the inverse of $[f(t)]$.

and the virtue of passing to homotopy classes is that now the product (12.3) is in fact associative. The proof is in Figure 80. Written out in excruciating detail the homotopy is

$$F(s, t) = \begin{cases} f_1(\frac{4}{s+1}t) & 0 \leq t \leq \frac{s+1}{4} \\ f_2(4t - (s+1)) & \frac{s+1}{4} \leq t \leq \frac{s+2}{4} \\ f_3(\frac{4}{2-s}(t - \frac{s+2}{4})) & \frac{s+2}{4} \leq t \leq 1 \end{cases} \quad (12.4)$$

4. Since we have an associative product on $\pi_1(X, x_0)$ we are now ready to define a group structure. The identity element is clearly given by the (homotopy class of the) constant loop: $f(t) = x_0$. If a homotopy class is represented by a loop $f(t)$ then the inverse is represented by running the loop backwards: $g(t) := f(1 - t)$. The two are joined at $t = 1/2$, and since this is in the open interval $(0, 1)$ the image can be deformed away from x_0 . See Figure 81. *Thus, with the group operation defined by concatenation in the sense of (12.3) the set of homotopy classes $\pi_1(X, x_0)$ is a group. It is known as the fundamental group based at x_0 .*
5. *Basepoint dependence:* If X is path connected then if we choose a different basepoint x'_0 we can define an isomorphism between homotopy groups by choosing a path γ from x_0 to x'_0 and mapping $[f] \mapsto [\bar{\gamma} \star f \star \gamma]$ where $\bar{\gamma}$ is $\gamma(1 - t)$. Let us call this isomorphism $\psi_\gamma : \pi_1(X, x_0) \rightarrow \pi_1(X, x'_0)$. Of course, ψ_γ only depends on the homotopy class (with fixed endpoints) of γ . Therefore, to each (pathwise) connected component of X we can assign a homotopy group, defined up to isomorphism. But to define an actual group, and not just an isomorphism class of a group, one must *choose* a basepoint x_0 . Moreover, the isomorphism between $\pi_1(X, x_0)$ and $\pi_1(X, x'_0)$ *is not canonical*. What this means is that if we choose γ_1 and γ_2 in different homotopy classes (with fixed endpoints) to define our isomorphisms then ψ_{γ_1} and ψ_{γ_2} will be “different.” Indeed the composition $\psi_{\bar{\gamma}_2} \circ \psi_{\gamma_1} : \pi_1(X, x_0) \rightarrow \pi_1(X, x_0)$ will be the inner automorphism of $\pi_1(X, x_0)$ defined by conjugation with the homotopy class of the closed loop at x_0

given by $\gamma_2 \star \bar{\gamma}_1$. To see this note that

$$\begin{aligned}\psi_{\bar{\gamma}_2} \circ \psi_{\gamma_1}([f]) &= [\gamma_2 \star (\bar{\gamma}_1 \star f \star \gamma_1) \star \bar{\gamma}_2] \\ &= [(\gamma_2 \star \bar{\gamma}_1) \star f \star (\gamma_1 \star \bar{\gamma}_2)] \\ &= h \cdot [f] \cdot h^{-1}\end{aligned}\tag{12.5}$$

where $h = [\gamma_2 \star \bar{\gamma}_1]$. If X is connected I will sometimes be sloppy and omit the basepoint and just write $\pi_1(X)$. But this is just shorthand, and it can lead you astray! For example, this basepoint dependence will be quite important when discussing the Galois correspondence between coverings of X and subgroups of $\pi_1(X)$.

6. A connected space such that $\pi_1(X, x_0)$ is the trivial group is called *simply connected*.
7. If $F : X \rightarrow Z$ takes $x_0 \in X$ to $z_0 \in Z$ then we defined above $F_* : \pi_1(X, x_0) \rightarrow \pi_1(Z, z_0)$. This can be shown to be a group homomorphism. In particular, if F is a homotopy equivalence, then it is a group isomorphism.
8. In algebraic topology books a major result which is proved is the *Seifert-van Kampen theorem*. This can be useful because it allows one to compute $\pi_1(X, x_0)$ by breaking up X into simpler pieces. Specifically, suppose that $X = U^+ \cup U^-$ is a union of two open path-connected subsets and that $U^{+-} := U^+ \cap U^-$ is also path-connected and contains x_0 . Now suppose we know presentations of the fundamental groups of the pieces U^+, U^-, U^{+-} in terms of generators and relations:

$$\begin{aligned}\pi_1(U^+, x_0) &\cong \langle g_1^+, \dots, g_{n^+}^+ | R_1^+, \dots, R_{m^+}^+ \rangle \\ \pi_1(U^-, x_0) &\cong \langle g_1^-, \dots, g_{n^-}^- | R_1^-, \dots, R_{m^-}^- \rangle \\ \pi_1(U^{+-}, x_0) &\cong \langle g_1^{+-}, \dots, g_{n^{+-}}^{+-} | R_1^{+-}, \dots, R_{m^{+-}}^{+-} \rangle\end{aligned}\tag{12.6}$$

Then the recipe for computing $\pi_1(X, x_0)$ is this: Denote the injection $\iota^+ : U^{+-} \rightarrow U^+$ and $\iota^- : U^{+-} \rightarrow U^-$. Then the generators of $\pi_1(U^{+-}, x_0)$ push forward to words in g_i^+ or g_i^- , respectively:

$$\begin{aligned}\iota_*^+(g_i^{+-}) &:= W_i^+ & i = 1, \dots, n^{+-} \\ \iota_*^-(g_i^{+-}) &:= W_i^- & i = 1, \dots, n^{+-}\end{aligned}\tag{12.7}$$

Finally, we have the presentation:

$$\pi_1(X, x_0) \cong \langle g_1^+, \dots, g_{n^+}^+, g_1^-, \dots, g_{n^-}^- | R_\alpha \rangle\tag{12.8}$$

where the relations R_α include the *old relations*

$$R_1^+, \dots, R_{m^+}^+, R_1^-, \dots, R_{m^-}^-\tag{12.9}$$

and a set of *new relations*:

$$W_1^+(W_1^-)^{-1}, \dots, W_{n^{+-}}^+(W_{n^{+-}}^-)^{-1}\tag{12.10}$$

It is obvious that these are relations on the generators. What is not obvious is that these are the only ones. Note that in the final presentation the generators g_i^{+-} and the relations R_i^{+-} have dropped out of the description.

9. There is a beautiful generalization of the fundamental group of a path connected space X known as the *fundamental groupoid* of X . This is not a group, but a category. The “objects” are the points of X . The “morphisms” from x_0 to x_1 are the homotopy classes of continuous paths from x_0 to x_1 . A groupoid is a category in which all morphisms are invertible: This corresponds to running the path backwards. The set of morphisms of an object to itself $\text{Hom}(x_0, x_0)$ forms a group: This is the fundamental group based at x_0 .

Examples

1. $\pi_1(S^1, x_0)$ is a free group on one generator, and hence isomorphic to $\mathbb{Z}$. A representative path of the generator can be taken to be $f(t) = \exp[2\pi it]$, where we represent the target S^1 as the unit circle in the complex plane. Another generator is the homotopy class of $g(t) = \exp[-2\pi it]$. This will be proven in Section **** below on covering space theory.
2. $\pi_1(S^1 \vee S^1)$ is the free group on two generators. This follows from the Seifert-van Kampen theorem and shows that, in general, π_1 can be a nonabelian group.
3. $\pi_1(S^2 - \{x_1, x_2, x_3\}; x_0)$ is the free group on two generators. Indeed, the three-punctured sphere can be deformation retracted to the wedge of two circles.

Exercise

Let $f(t) = \exp[2\pi it]$ be a generator of $\pi_1(S^1, x_0)$ with $x_0 = 1$. Show that

$$\underbrace{f \star \cdots \star f}_{n \text{ times}} \tag{12.11}$$

is the loop $f_n(t) = \exp[2\pi int]$.

Exercise

What is the fundamental group of the wedge of g copies of the circle?

12.2.1 Remark on winding number

The result $\pi_1(S^1, x_0) \cong \mathbb{Z}$ is related to some simple facts about analytic functions on the plane.

Let $f(z)$ be an analytic function in the neighborhood of the unit circle in the complex plane. Suppose that $f(z) \neq 0$ in this neighborhood. Then we can unambiguously define

the phase of f by $f(z) = |f(z)|\Phi(z)$ where $|\Phi(z)| = 1$. The map from the unit circle $|z| = 1$ to the unit circle given by $z \mapsto \Phi(z)$ has a well-defined homotopy class, and the associated winding number is called the degree of f : It can be written

$$\deg(f) := \oint \frac{f'(z)}{f(z)} \frac{dz}{2\pi i} = \frac{1}{2\pi i} \oint d\log f(z) \quad (12.12)$$

If f has a nonzero degree then there is no single-valued logarithm of f in a neighborhood of the unit circle. Only when the degree is zero is there a well-defined real-valued (real analytic) function $\theta(z, \bar{z})$ such that $\Phi(z) = e^{2\pi i \theta(z, \bar{z})}$. In this case we have a single valued function $\log f(z)$.

A nice mathematical corollary of this connection is the fundamental theorem of algebra: Every nonconstant polynomial has a complex root. We can prove this by contradiction: Let $p(z)$ be the hypothetical nonconstant polynomial with no root. Then the phase $\Phi(z)$ is well-defined on the whole plane. Moreover, the winding number is zero:

$$\deg p = \oint \frac{p'(z)}{p(z)} \frac{dz}{2\pi i} = 0 \quad (12.13)$$

If $p(z)$ has no zeroes then by Cauchy's theorem we can shrink the contour to a point without changing the integral, and the answer clearly has to be zero. On the other hand, by Cauchy's theorem we can also deform the contour to large values of z , say on a contour $z(t) = Re^{2\pi i t}$ with $R \rightarrow \infty$. WLOG we can assume that $p(z) = a_n z^n + a_{n-1} z^{n-1} + \dots + a_0$ with $n > 0$ and $a_n \neq 0$. At large radius the z^n term dominates, and since $\deg(p)$ must be an integer we must have

$$\oint \frac{p'(z)}{p(z)} \frac{dz}{2\pi i} = \oint \frac{n z^{n-1}}{z^n} \frac{dz}{2\pi i} = n \quad (12.14)$$

This is a contradiction, so $p(z)$ must have a root.

Exercise

Show that every nonconstant quaternion-valued polynomial on the quaternions has a quaternionic root.

12.2.2 Surface groups

Using the construction of closed surfaces as polygons with identifications discussed in Section *** above it is easy to give a presentation of the fundamental groups of surfaces. Note that if we cut a disk out of the polygon then there is a homotopy equivalence with a wedge product of a collection of circles. Therefore, we have an obvious set of generators with one relation: The closed path described by circling the boundary of the polygon must be trivial. Applying this to the three basic cases we get:

1. If S is S^2 then it is simply connected.

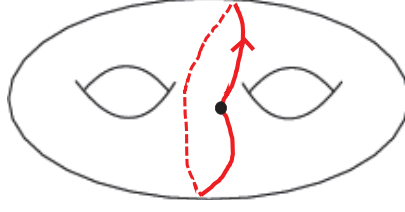


Figure 82: Is the homotopy class of this curve trivial? Is it trivial in the first homology group?

2. If S is an orientable genus g surface with p boundaries then

$$\pi_1(S, x_0) = \langle a_i, b_i, c_s \mid \prod_{i=1}^g [a_i, b_i] \prod_{s=1}^p c_s = 1 \rangle \quad (12.15)$$

3. If S is an unorientable surface with q Mobius strips and p boundaries then

$$\pi_1(S, x_0) = \langle \gamma_i, c_s \mid \prod_{i=1}^q \gamma_i^2 \prod_{s=1}^p c_s = 1 \rangle \quad (12.16)$$

Remarks:

1. One can, of course, consider noncompact surfaces, but these are more involved, since they can have infinite numbers of handles.
2. The above presentations are rigorously derived from the Seifert-van Kampen theorem.
3. In many applications to string theory we often speak of *punctures* rather than boundaries. If we remove a point from a surface then the resulting space is noncompact. It has the same fundamental group as the space obtained by replacing the puncture by a small hole, so that the surface has a boundary (and its closure is compact, with boundary).

Exercise

- a.) Show that $\pi_1(\mathbb{RP}^2) \cong \mathbb{Z}_2$.

- b.) Now consider $\mathbb{RP}^2$ as $S^2/\sim$ where the equivalence relation identifies antipodal points. Identify a simple curve representing a nontrivial element and explain why it squares to zero.
- c.) Show that if S is the Mobius band or the cylinder then $\pi_1(S) \cong \mathbb{Z}$.
- d.) Using the presentation above show that $\pi_1(T^2) \cong \mathbb{Z} \oplus \mathbb{Z}$.
- e.) Are any other surface groups abelian?
- f.) Do any other surface groups have nontrivial torsion subgroups?
-

Exercise *Abelianization of the fundamental groups of closed surfaces*

The *abelianization* of a group G is the quotient of the commutator subgroup N generated by all group commutators $[g_1, g_2] := g_1 g_2 g_1^{-1} g_2^{-1}$, for all pairs of elements of G .

- a.) Show N is a normal subgroup and that G/N is an abelian group.
- b.) Compute the abelianization of the fundamental groups of closed surfaces. This is a finitely generated abelian group. Give its rank and torsion subgroup.

Remark: One can show that if X is connected the abelianization of $\pi_1(X, x_0)$ is isomorphic to the first homology group $H_1(X; \mathbb{Z})$.

- c.) Consider the based curve shown in 82. Does it define a trivial element of $\pi_1(X, x_0)$? Does it define a trivial element of $H_1(X; \mathbb{Z})$?
-

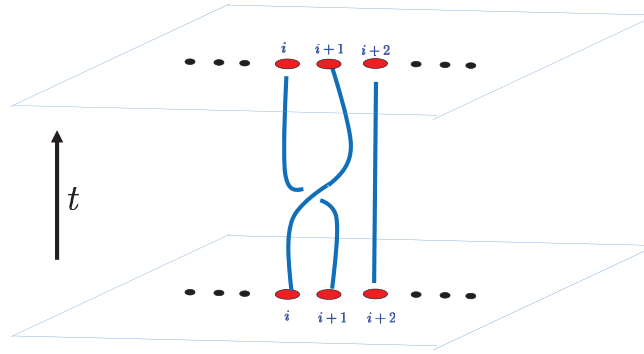


Figure 83: The time development shown here, with $0 \leq t \leq 1$, represents a *closed* path in $C_n(\mathbb{R}^2)$. It represents the generator $\tilde{\sigma}_i$ mentioned in Chapter 5.1.

12.2.3 Braid groups

The braid group was defined in terms of generators and relations in Section 5.1 of Chapter 1. It can be interpreted as a fundamental group where we consider the configuration space

of n points in the plane.

To define a *configuration space* take any space X and consider $X^n - \Delta$ where Δ is the space of tuples $(x_1, \dots, x_n)$ where $x_i = x_j$ for some pair of distinct indices i, j . We now consider the quotient space where we consider $(x_1, \dots, x_n) \sim (y_1, \dots, y_n)$ if there is a permutation $\sigma \in S_n$ so that $x_i = y_{\sigma(i)}$. Then

$$C_n(X) := (X^n - \Delta) / \sim \quad (12.17)$$

is the configuration space of n unordered points in X .

In the physics of identical particles, this would be the space of configurations of n particles in a space X . The justification for eliminating Δ could arise if the particles have hard cores or repulsive forces.

For example, in $2 + 1$ dimensional physics with identical particles, we could consider $C_n(\mathbb{R}^2)$. Then the braid group can be defined as

$$\mathcal{B}_n := \pi_1(C_n(\mathbb{R}^2), [\vec{x}_0]) \quad (12.18)$$

The figures in Section 5.1 give a pictorial justification for the relations on the generators. We could, for example consider $\vec{x}_0$ to be a line of points along, say, the x -axis. Then a simple braiding as shown in Figure 83 defines a closed loop in $C_n(\mathbb{R}^2)$ and it can be shown that the homotopy classes of these closed loops generate $\mathcal{B}_n$. Moreover, it can be shown that the only relations are the geometrically obvious ones:

$$\begin{aligned} \tilde{\sigma}_i \tilde{\sigma}_j &= \tilde{\sigma}_j \tilde{\sigma}_i & |i - j| \geq 2 \\ \tilde{\sigma}_i \tilde{\sigma}_{i+1} \tilde{\sigma}_i &= \tilde{\sigma}_{i+1} \tilde{\sigma}_i \tilde{\sigma}_{i+1} \end{aligned} \quad (12.19)$$

Exercise

If we consider $f : \mathbb{R}^2 \rightarrow S^2$ then we have a homomorphism $f_* : \mathcal{B}_n \rightarrow \pi_1(C_n(S^2))$. It can be shown that the image of the generators $\tilde{\sigma}_i$ (which we denote $\hat{\sigma}_i$) now have two new relations:

$$\hat{\sigma}_1 \cdots \hat{\sigma}_{n-2} \hat{\sigma}_{n-1}^2 \hat{\sigma}_{n-2} \cdots \hat{\sigma}_1 = 1 \quad (12.20)$$

$$(\hat{\sigma}_1 \cdots \hat{\sigma}_{n-1})^n = 1 \quad (12.21)$$

Draw pictures of the braids corresponding to (12.20) and (12.21) and explain why they should hold.

Exercise

Construct a mapping (natural up to homotopy):

$$C_k(\mathbb{R}^2) \times \left(C_{p_1}(\mathbb{R}^2) \times \cdots \times C_{p_k}(\mathbb{R}^2) \right) \rightarrow C_{p_1 + \cdots + p_k}(\mathbb{R}^2)$$

The existence of this system of maps leads to the subject of *operads*.

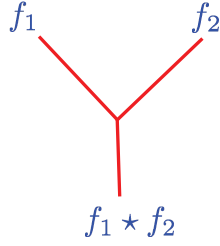


Figure 84: The composition of two closed loops is depicted by a trivalent vertex.

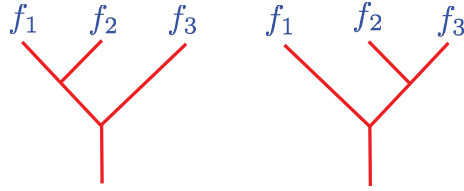


Figure 85: The two ways of composing three loops can be represented by two different rooted trees with trivalent vertices.

12.2.4 Digression: A_∞ spaces

Warning: This is a more advanced topic. You might wish to skip this section.

We noted above that for three closed loops in X based at x_0 , in general $f_1 \star (f_2 \star f_3)$ is *not* the same loop as $(f_1 \star f_2) \star f_3$. When defining the fundamental group we forced “multiplication of loops” to be associative by passing to the equivalence class under homotopy. Actually, the fact that the loops are not equal, but only homotopic is the tip of the iceberg of an interesting mathematical structure which plays some role in modern physical mathematics. In this section we sketch a bit of that structure.

Given a *pointed space* (X, x_0) , we can define the *based loop space*:

$$\Omega_* X := \{\gamma : (S^1, s_0) \rightarrow (X, x_0)\} \quad (12.22)$$

Note that ΩX is itself a pointed space, with the basepoint being the constant loop which takes all of S^1 to x_0 . We can also define it as:

$$\Omega_* X := \{f : [0, 1] \rightarrow X \quad \text{with} \quad f(0) = f(1) = x_0\} \quad (12.23)$$

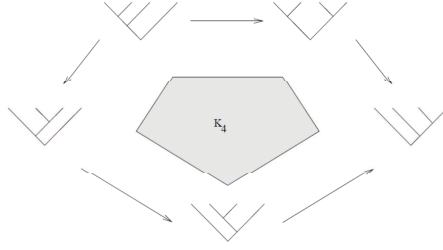


Figure 86: The five ways of composing four loops can be represented by five different rooted trees with trivalent vertices. These form the vertices of a pentagon K_4 . The figure is taken from B. Keller, “Introduction to A_∞ algebras and modules,” arXiv:math/9910179.

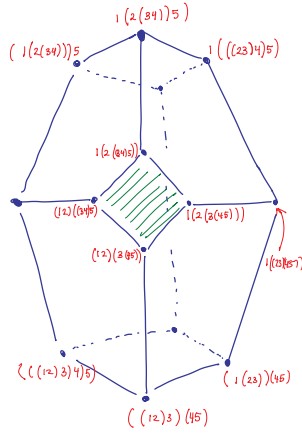


Figure 87: The 14 ways of composing f_1, f_2, f_3, f_4, f_5 , in that order, can be viewed as vertices of a polytope. The two vertices are joined by an edge if the two products are related by a basic associativity map. The resulting polytope has faces consisting of pentagons and squares.

The concatenation of loops defined in (12.2) defines a product

$$m_2 : \Omega_* X \times \Omega_* X \rightarrow \Omega_* X. \quad (12.24)$$

We will find it very useful to have a pictorial representation of products of loops in terms of trees. The basic multiplication is shown in Figure 84.

Now, in Figure 80 we gave a homotopy between

$$(f_1 \star f_2) \star f_3 = m_2(m_2(f_1, f_2), f_3) \quad (12.25)$$

and

$$f_1 \star (f_2 \star f_3) = m_2(f_1, m_2(f_2, f_3)) \quad (12.26)$$

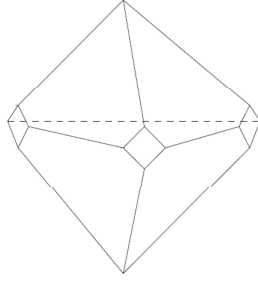


Figure 88: The 14 ways of composing five loops can be represented by different rooted trees with trivalent vertices. They form the vertices of the fifth associahedron. The figure is taken from B. Keller, “Introduction to A_∞ algebras and modules,” arXiv:math/9910179.

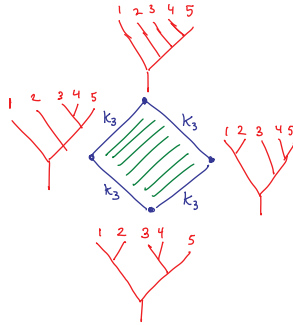


Figure 89: The square faces of the fifth associahedron correspond to polytopes of the form $K_3 \times K_3$ and hence it is easy to extend the map from the edges to the face.

The existence of this homotopy can be interpreted as the statement that there is a map

$$m_3 : K_3 \times \Omega_* X \times \Omega_* X \times \Omega_* X \rightarrow \Omega_* X \quad (12.27)$$

where $K_3 := [0, 1]$ is the interval. The value of

$$m_3(s, f_1, f_2, f_3) \quad (12.28)$$

where $s \in [0, 1]$ and f_1, f_2, f_3 are three loops at x_0 is the based loop in X defined on the RHS of (12.4). The values of this map at the boundaries of K_3 , that is, at $s = 0$ and $s = 1$ give the two possible compositions involving m_2 . The restriction of the map to the two boundaries of K_3 can be depicted as in Figure 85. We will call m_3 an *associativity map*.

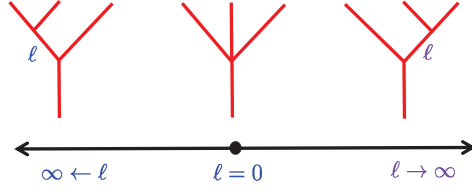


Figure 90: To construct K_3 from the moduli spaces of trees we compactify the two cells $(0, \infty)$ for the two topologies of tree to $[0, \infty)$ and identify the $\ell = 0$ points, as shown. Then we use a homeomorphism to make the semi-infinite intervals finite, and/or add a point at infinity. Thus, for example, we could identify $x \in [0, 1]$ with a length $\ell = \frac{x}{e^{1-x} - 1}$. Thus, we can compactify by letting edge lengths be valued in $\ell \in [0, \infty]$.

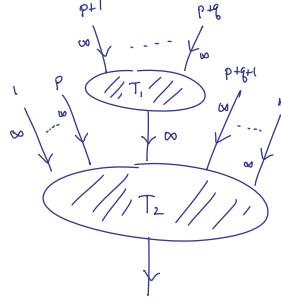


Figure 91: If we regard K_n as the compactified moduli space of ordered metric trees then the $(n - 3)$ -dimensional faces are associated with trees where precisely one internal edge has gone to infinite length.

Now, let us ask what happens when we compose four maps f_1, f_2, f_3, f_4 , in that order.

There are now five ways to do this:

$$\begin{array}{ccc}
 & ((f_1 \star f_2) \star f_3) \star f_4 & \longrightarrow (f_1 \star f_2) \star (f_3 \star f_4) \\
 & \swarrow & \searrow \\
 (f_1 \star (f_2 \star f_3)) \star f_4 & & f_1 \star (f_2 \star (f_3 \star f_4)) \\
 & \searrow & \swarrow \\
 & f_1 \star ((f_2 \star f_3) \star f_4) &
 \end{array}
 \tag{12.29}$$

The arrows represent the basic homotopy expressed by the existence of m_3 . We can view each of the corresponding multiplications in terms of trees, and think of them as attached to the vertices of a pentagon, as in Figure 86.

We have associated each way of putting parentheses to a vertex of a pentagon. We can define a map

$$m_4 : \partial K_4 \times (\Omega_* X)^4 \rightarrow \Omega_* X \tag{12.30}$$

where ∂K_4 is the *boundary* of the pentagon. To do this we think of each edge of the boundary as a copy of K_3 and we use a map of type m_3 defined in (12.31), applied to the appropriate triple of maps, on each segment of the boundary ∂K_4 .

Now comes a key new point: It turns out one can *fill in* the map from the boundary of the pentagon to a map from the entire pentagon. Let K_4 denote the filled in pentagon. We are claiming that the map of equation (12.30) can be continuously extended to a single map

$$m_4 : K_4 \times \Omega_* X \times \Omega_* X \times \Omega_* X \rightarrow \Omega_* X \tag{12.31}$$

We call m_4 a *higher associativity map*. The proof that it exists is an even more tedious formula than (12.4).

Next, we can consider the multiplication of 5 loops, $f_1, \dots, f_5$. There are 14 ways to do this. We define vertices of a similar polytope, denoted by K_5 and illustrated in Figure 87 and Figure 88. We can define a map $m_5 : E \times (\Omega_* X)^5 \rightarrow \Omega_* X$ on the edges using m_3 . We now want to extend this map to the faces. On the faces which are pentagons we can extend it using m_4 . Some faces are squares which we can view as $K_3 \times K_3$. We can easily extend m_5 across these squares because the two associativities are independent. See, for example, Figure 89. Now, the next nontrivial claim is that the map on ∂K_5 can be extended to a map:

$$K_5 \times (\Omega_* X)^5 \rightarrow \Omega_* X \tag{12.32}$$

which has the property that, when restricted to the boundaries of K_5 the map becomes one of the extended associativity maps we have already met.

Stasheff constructed a sequence of polytopes K_n of dimension $(n - 2)$ together with maps

$$K_n \times (\Omega_* X)^n \rightarrow \Omega_* X \tag{12.33}$$

which have the kind of compatibility we mentioned above: The restriction of the map to the boundary ∂K_n is a collection of the higher associativity laws we found above. By

counting trees one can show that the polytope K_{n+1} has a number of vertices equal to the Catalan number:

$$C_n = \frac{1}{n+1} \binom{2n}{n} = \binom{2n}{n} - \binom{2n}{n+1} = \prod_{j=2}^n \frac{n+j}{j} \quad (12.34)$$

One way of constructing the polytopes K_n is to look at all trees with n ordered inputs and one output. The trees have trivalent vertices and are directed downward, as in the above figures. Then we assign positive lengths $\ell \in (0, \infty)$ to the internal edges. Thus, for each tree we associate a moduli space of such trees that is equivalent to an open cell $(0, \infty)^{n-2}$. Now, we can joint together these cells by letting some lengths go to zero. Thus gluing together some of the cells at some of their boundaries. See Figure 90

Now, compactifying the ends where some lengths go to ∞ gives the polytope K_n , known as the *associahedron*. The boundary of K_n is made up of lower associahedra. The $(n-3)$ -dimensional faces are associated with trees where one internal length has gone to infinity. See Figure 91. Thus the boundaries are associated with $K_q \times K_{n-q+1}$, and m_n restricted to these faces is - roughly speaking - of the form

$$m_{n-q+1}(f_1, \dots, f_p, m_q(f_{p+1}, \dots, f_{p+q}), f_{p+q+1}, \dots, f_n) \quad (12.35)$$

This leads to a definition

Definition: An A_∞ space $\mathfrak{X}$ is a topological space such that there exists a sequence of maps m_n , $n \geq 2$

$$m_n : K_n \times \mathfrak{X}^n \rightarrow \mathfrak{X} \quad (12.36)$$

(where $K_2 = pt$) which are compatible when restricted to ∂K_n , that is, of the form (12.35).

The main theorem of the subject says that if $\mathfrak{X}$ is an A_∞ -space then there exists another topological space X and a homotopy equivalence between $\mathfrak{X}$ and $\Omega_* X$.⁷⁰

Some sources: Most of the material on A_∞ is very algebraic. I have followed the discussions in

1. B. Keller, “Introduction to A_∞ algebras and modules,” arXiv:math/9910179.
2. B. Keller, ...
3. P.S. Aspinwall et. al., *Dirichlet Branes and Mirror Symmetry*, AMS Clay mathematics monographs, vol. 4, 2009

Remarks:

1. The associahedron comes up in several other related contexts. Whenever we want to multiply formal noncommutative and potentially non-associative variables we can enumerate the ways of multiplying n objects, $x_1, x_2, \dots, x_n$ (in that order!) to the rooted trivalent trees as above.

⁷⁰Reference: J.F. Adams, *Infinite Loop Spaces*, Annals of Math Studies, **90**, Princeton Univ. Press, Section 2.2.

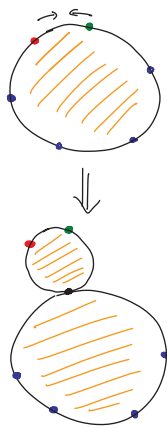


Figure 92: The “Mickey Mouse” compactification of the moduli space of disks with marked boundary points. This is a special case of a deep result of Deligne and Mumford on the compactification of the moduli spaces of curves.

2. The number of triangulations of a plane polygon with $(n + 2)$ vertices is again the Catalan number. (You can map this to binary rooted trees with ordered inputs.) Flipping an edge of a triangulation gives another one, and this defines the 1-simplices of K_{n+1} .
3. Another appearance of the associahedron is in Riemann surface theory. A Riemann surface has more structure than just a topological surface: It has a metric up to conformal rescaling, or, equivalently, a complex structure. If we consider the space of equivalence classes of the Riemann surface given by a disk with n ordered points on the boundary we are naturally led to consider K_n . The proper compactification is not obvious and the main idea is illustrated in Figure 92. This is one way A_∞ algebras enter into conformal field theory, string theory, string field theory, and topological field theory.⁷¹

Exercise

Show that the Catalan numbers can be defined recursively by $C_0 = 1$ and

$$C_{n+1} = \sum_{j=0}^n C_j C_{n-j} \quad (12.37)$$

Interpret this in terms of counting of ordered rooted binary trees.

⁷¹For one representative paper on the subject see, M. R. Gaberdiel and B. Zwiebach, “Tensor constructions of open string theories. 1: Foundations,” Nucl. Phys. B **505**, 569 (1997) [hep-th/9705038].

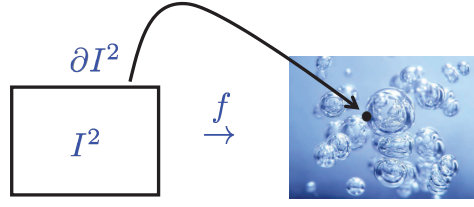


Figure 93: To define $\pi_2(X, x_0)$ we consider maps from I^2 which map the boundary ∂I^2 (the black square on the left) into a single point $x_0 \in X$, represented by the black point on the right. The interior of I^2 maps into the space in some continuous manner. For example, it might wrap around the soap bubble on the right.

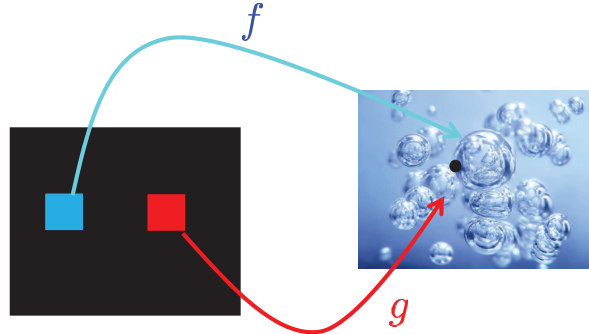


Figure 94: Combining two maps of the square to define a product map $f * g : (I^n, \partial I^n) \rightarrow (X, p_0)$. We draw the case $n = 2$. All of the black region maps to the black point x_0 on the right. The maps f, g define the maps from the smaller squares. There are clearly many choices made here but the homotopy class $[f * g]$ only depends on $[f]$ and $[g]$.

12.3 Higher homotopy groups

Choosing a point q_0 on an n -dimensional sphere S^n higher fundamental groups are, as a set:

$$\pi_n(X, p_0) := [(S^n, q_0), (X, p_0)] = [S^n, X]_* \quad (12.38)$$

Alternatively, if I^n is the n -cube and ∂I^n is the boundary then the n^{th} homotopy group is:

$$\pi_n(X, p_0) = [(I^n, \partial I^n), (X, p_0)] \quad (12.39)$$

This is illustrated in 93.



Figure 95: Pictorial proof that $\pi_2(X)$ is abelian.

The product is best described in pictures: The product of the homotopy classes $[f]$ and $[g]$ of two maps is best defined by using a picture as in 94. We homotope the two maps so that we may take f, g to be constant outside of a little cube, necessarily mapping to the basepoint x_0 . Then we can insert the two little cubes inside a bigger cube to produce a new such map as in 94. If you insist on a precise formula we could take:

$$(f \star g)(t_1, t_2, \dots, t_n) := \begin{cases} f(2t_1, t_2, \dots, t_n) & 0 \leq t_1 \leq \frac{1}{2} \\ g(2t_1 - 1, t_2, \dots, t_n) & \frac{1}{2} \leq t_1 \leq 1 \end{cases} \quad (12.40)$$

and then $[f] \cdot [g] := [f \star g]$ is a definite multiplication rule. Of course, we could have defined $\star$ by concatenation along one of the other t_i -axes, or in many other ways.

Theorem. For $n > 1$ $\pi_n(X, p_0)$ are *abelian* groups.

Proof: This is natural because there is no natural order for concatenating maps from I^n for $n > 1$ (a consequence of the fact that S^{n-1} is connected, for $n > 1$).

Again, it is best explained with a picture as shown in 95. ♠

In general it is not easy to compute homotopy groups. We will illustrate a few techniques from which one can extract some information in Sections ****. We will just quote a few results.

Examples

1. For the circle we have

$$\pi_k(S^1, p_0) = \begin{cases} \mathbb{Z} & k = 1 \\ 0 & \text{else} \end{cases} \quad (12.41)$$

This will follow easily from covering space theory. See section *** below.

2. For a surface $\pi_k(S, p_0) = 0$ for $k > 1$. Again, using a standard representation of surfaces, this follows from covering space theory.
3. For an n dimensional sphere with $n > 1$ we have

$$\pi_k(S^n, p_0) = 0 \quad 1 \leq k < n \quad (12.42)$$

Here is an heuristic argument (which can be made completely rigorous): First, an homotopy class $[f]$ has a representative by a *differentiable* map $f : S^k \rightarrow S^n$. (This is plausible but not trivial and is a special case of Proposition 17.8 of Bott and Tu.) Second, for $k < n$ the image of a differentiable map f cannot be surjective. Again, this is very intuitive, but not trivial. It is a consequence of Sard's theorem. Once we have a representative which is not onto the remainder of the argument is easy: f maps S^k to a subspace of $S^n - \{Q\}$ for some point Q , but that is homeomorphic (by stereographic projection) to $\mathbb{R}^n$, and hence contractible. We also have

$$\pi_n(S^n, p_0) \cong \mathbb{Z} \quad (12.43)$$

We will prove this in Section **** below.

4. The higher homotopy groups of spheres, that is, $\pi_{n+k}(S^n, p_0)$ for $n > 1$ and $k > 0$ is a mathematical object that remains a topic of central interest in modern algebraic topology. Many interesting facts are known. As a small sampling: For large n , $\pi_{n+k}(S^n, p_0)$ becomes independent of n . To be precise, they become independent of n for $n \geq k + 2$. These are called the *stable* homotopy groups of spheres π_k^S . The higher homotopy groups (stable and unstable) are finite groups except for $\pi_{4n-1}(S^{2n})$ which contains an infinite cyclic component. We will see that $\pi_3(S^2) = \mathbb{Z}$, a case of great importance in physics. As an example, consider $\pi_{n+1}(S^n)$ $n = 2$ is in the unstable range, and $n \geq 3$ is the stable range, and $\pi_{n+1}(S^n) \cong \mathbb{Z}_2$ for $n \geq 3$. Tables of homotopy groups of spheres are available, for example, on the Wikipedia article.

Exercise Homotopy of product spaces

a.) Show that

$$\pi_q(X \times Y) \cong \pi_q(X) \times \pi_q(Y) \quad (12.44)$$

where on the RHS we have a direct product of groups. ⁷²

Warning: This is *not* true for cohomology and homology groups in general!

b.) Let T^n be the n -dimensional torus. What is $\pi_k(T^n)$ for $k > 1$?

c.) Are there nontrivial homotopy classes of maps $T^2 \rightarrow S^2$?

⁷²Hint: Show that every map S^q into $X \times Y$ is of the form $x \rightarrow (f(x), g(x))$.

Exercise *Homotopy groups of groups*

If G is a group then $\pi_n(G, e)$ has some special features:

- a.) Show that if $f_i : (I^n, \partial I^n) \rightarrow (G, e)$ are two maps then the map $f_1 \cdot f_2$ defined by the group product on the target is homotopic to the homotopy product of f_1 and f_2 .
 - b.) In particular show that $\pi_1(G, e)$ is abelian.
-

12.4 Homotopy groups and homotopy equivalence

Now we can ask how homotopy groups of different spaces are related. If $F : X \rightarrow Y$ is a map between different spaces then to any $f : (S^n, q_0) \rightarrow (X, p_0)$ we can associate $F \circ f : (S^n, q_0) \rightarrow (Y, F(p_0))$. Moreover, this association respects homotopic equivalence so we get a map

$$F_* : \pi_n(X, p_0) \rightarrow \pi_n(Y, F(p_0)) \quad (12.45)$$

defined by $F_*([f]) := [F \circ f]$. Furthermore, a simple exercise shows that F_* is a group homomorphism.

Theorem If two spaces are homotopically equivalent then they have isomorphic homotopy groups.

Proof: If $F : X \rightarrow Y$ has homotopy inverse $G : Y \rightarrow X$ then $F_* : \pi_n(X) \rightarrow \pi_n(Y)$ is a group isomorphism with inverse G_* . ♠

Remarks

1. Thus, the groups $\pi_n(X)$ are topological invariants of X . Since spaces can be homotopy equivalent but not homeomorphic they will not always distinguish different topological spaces. For example $\mathbb{R}^n$ and a point have the same homotopy groups. In two dimensions π_1 distinguishes the different compact surfaces. But the homotopy groups can fail to distinguish homeomorphism types of compact spaces already in dimension three. Some lens spaces can have isomorphic π_n , and yet not be homeomorphic. (For details see Massey's book.)
2. We can now finally answer some questions we raised in Section §??: S^n cannot be homotopy equivalent to S^m for $n \neq m$. Thus $\mathbb{R}^n$ cannot be homeomorphic to $\mathbb{R}^m$ for $n \neq m$, for if they were the one-point compactifications would be homeomorphic.
3. What about the converse to the above theorem? That is, if X and Y have the same homotopy groups are they homotopy equivalent? Strictly speaking the converse is false. A simple counterexample is $S^2 \times \mathbb{R}P^3$ and $\mathbb{R}P^2 \times S^3$. Both are obtained by taking a quotient of $S^2 \times S^3$ by an antipodal map. On the one hand, using covering space theory of Section §13 it follows that they have the same set of homotopy groups. On the other hand, one can show that if two topological manifolds are homotopy equivalent then they are both orientable or both unorientable.

4. Nevertheless, if there is a *concrete map* $f : X \rightarrow Y$ which induces an isomorphism of the homotopy groups then one *can* conclude that the map is a homotopy equivalence. This is the famous Whitehead theorem:

Theorem. A map $f : X \rightarrow Y$ is a homotopy equivalence iff f_* is an isomorphism on all π_n .

See Bredon, Corollary 11.14. It applies to reasonable spaces X, Y . For example, the CW complexes described in chapter *** will satisfy the conditions.

12.4.1 Homotopy Invariants of maps between spaces

A frequently asked question is: Given two maps $f_1, f_0 : X \rightarrow Y$ are they homotopic?

This comes up in physics in many ways:

1. In field theory, when we classify topological sectors.
2. In string theory, we might ask when maps of the string worldsheet into spacetime are continuously related to each other.
3. In condensed matter physics, we can associate a continuous projection operator to the filled bands over the Brillouin zone and the homotopy class of that map is an invariant of an insulator.

In general it is a very difficult question to say whether or not f_0 and f_1 are homotopic.

Some important homotopy invariants of maps, which can distinguish different classes, are:

1. The group homomorphisms: $f_* : \pi_n(X, x_0) \rightarrow \pi_n(Y, f(x_0))$ for all $n \geq 1$.
2. The group homomorphisms $f_* : H_n(X; \mathbb{Z}) \rightarrow H_n(Y; \mathbb{Z})$
3. The group homomorphisms $f^* : H^n(Y; \mathbb{Z}) \rightarrow H^n(X; \mathbb{Z})$

The data 2,3 are not really independent, but the data 1,2 are really independent data. For example, taking $n = 3$ and $X = Y = S^2$, $H_3(S^2) = 0$ but $\pi_3(S^2) \cong \mathbb{Z}$ and f_* can be an interesting multiplication map. One set of maps can be trivial and the other nontrivial.

Homotopic maps induce the same homomorphism of groups, so, one quick test of whether two maps can be homotopic is whether the induced group homomorphisms f_* and f^* are the same.

12.5 Homotopy and its relation to loop space

Recall that the map from $\mathcal{C}(X \times Y, Z)$ to $\mathcal{C}(X, \mathcal{C}(Y, Z))$ is continuous. Using this we can prove that

Theorem: $\pi_{k-1}(\Omega_* X, \bar{x}_0) = \pi_k(X, x_0)$, for $k \geq 2$.

Here the basepoint $\bar{x}_0$ of $\Omega_* X$ used to compute the homotopy group is the constant loop at x_0 .

There is an interesting “adjoint” to the loop operation known as suspension. To define it we first define a more general construction, known as the *smash product*.

Recall that the one-point union, or wedge, of two pointed spaces X, Y is obtained by identifying *only* their basepoints. That is $X \vee Y = (X \amalg Y)/x_0 \sim y_0$. The *smash product* of two pointed spaces is defined by identifying all of $X \vee Y$ to a single basepoint:

$$X \wedge Y := (X \times Y)/X \vee Y \quad (12.46)$$

An especially important case is the *suspension*⁷³

$$SX := S^1 \wedge X = (X \times S^1)/(X \times \{s_0\} \cup \{x_0\} \times S^1) \quad (12.47)$$

where s_0 is a basepoint on S^1 .

Theorem: Consider the spheres to be pointed spaces. Then SS^{n-1} is homeomorphic to S^n for all $n \geq 1$.

You can convince yourself of this by drawing some simple pictures.

A nice result is

Theorem: $[SX, Y]_* = [X, \Omega Y]_*$.

To prove this simply write a pointed map $f : SX \rightarrow Y$ as $f(t, x)$ and then consider the loop γ_x at fixed x given by $\gamma_x : t \rightarrow f(t, x)$. Now on the RHS we identify this with the map $x \rightarrow \gamma_x \in \Omega Y$.

In this sense, suspension is “adjoint” to looping.

13. Fibrations and covering spaces

13.1 The lifting problem

Let us consider again the computation of $\pi_1(S^1, x_0)$. We think of S^1 as the unit circle in the complex plane and let us take $x_0 = 1$. A homotopy class would be represented by a function $f : [0, 1] \rightarrow S^1$ which we could try to write as $f(t) = \exp[2\pi i g(t)]$ where $f(0) = f(1) = 1$. We could surely take $g(0) = 0$, but then, if we continuously evolve $g(t)$, it might well be that $g(1) \neq 0$, rather, it could be some integer $g(1) = n \in \mathbb{Z}$. A homotopy of $f(t)$ would continuously change $g(t)$. In particular, a homotopy of f would continuously change $g(1)$. On the other hand, $g(1)$ is an integer. An integer which changes continuously must be constant. Thus, the integer $g(1)$ is only a function of the homotopy class of f . For example, if $f(t) = e^{2\pi i t}$ then $g(t) = t$. Note that $g(1) = 1 \neq g(0) = 0$. This shows, informally, that there is a (obviously surjective) homomorphism $\pi_1(S^1, x_0) \rightarrow \mathbb{Z}$. On the other hand, if $g(0) = 0$ then, we claim, there is no obstruction to smoothly deforming $g(t)$ to 0 throughout the interval. This shows - informally - that in fact $\pi_1(S^1, x_0) \cong \mathbb{Z}$.

⁷³Warning: There are two kinds of suspension in homotopy theory. This is called the “reduced suspension.”

We will now generalize the above reasoning and make it more rigorous. That will lead to some important mathematical ideas and constructions.

Here is the formal statement of the *lifting problem*:

Suppose we are given a continuous surjective map of topological spaces $p : E \rightarrow B$ and moreover we are given a map $\bar{f} : Y \rightarrow B$. Such a map is said to “have a lift f ” if there is a map $f : Y \rightarrow E$ with $p \circ f = \bar{f}$. In diagrams, given $p : E \rightarrow B$ and $\bar{f}$, “finding a lift of $\bar{f}$ ” means finding a function $f : Y \rightarrow E$ so that we have the commutative diagram:

$$\begin{array}{ccc} & E & \\ f \nearrow & \downarrow p & \\ Y & \xrightarrow{\bar{f}} & B \end{array} \quad (13.1)$$

Example: Suppose $E = \mathbb{R}$, $B = S^1$, and $p(x) = e^{2\pi i x}$. Then,

- If Y is a point and $\bar{f}(*) = z$, with $|z| = 1$ the lift $f(*)$ is simply some real number x so that $e^{2\pi i x} = z$. Of course, there are infinitely many such x ’s. Given one x_0 any other solution can be written as $x = x_0 + n$ for some $n \in \mathbb{Z}$. In this case, a lift of the map exists, and there are in fact infinitely many lifts.

- On the other hand, suppose now that $Y = S^1$ and we take $\bar{f} : S^1 \rightarrow S^1$ to be the identity map. Then, because there is no single valued function

$$\frac{1}{2\pi i} \log z \quad (13.2)$$

we cannot find a lift of $\bar{f}$. So, sometimes, there can be obstructions to lifting maps.

13.2 Homotopy lifting property

Suppose again we are given a continuous surjective map of topological spaces $p : E \rightarrow B$. Such a map is said to be a *fibration* if it satisfies a special property, called the *homotopy lifting property* or, equivalently, the *covering homotopy property*. It is defined as follows:

Suppose we are given:

- A map $f : Y \rightarrow E$, which is therefore a lift of $\bar{f} := p \circ f : Y \rightarrow B$.
- A homotopy $\bar{F}$ of $\bar{f}$. That is, a continuous family of maps $\bar{f}_t : Y \rightarrow B$ smoothly deforming $\bar{f}$. Equivalently, a continuous map

$$\bar{F} : [0, 1] \times Y \rightarrow B \quad (13.3)$$

such that $\bar{F}(0, y) = \bar{f}(y)$.

Then: $p : E \rightarrow B$ has the *covering homotopy property*, or has the *homotopy lifting property* if for any $Y, \bar{f}$ with a lift f there is a continuous map

$$F : [0, 1] \times Y \rightarrow E \quad (13.4)$$

with $p \circ F = \bar{F}$. That is, if, whenever we have $Y, \bar{f}, f$ as above there is a continuous family of maps f_t which lift $\bar{f}_t$ and with $f_0 = f$.

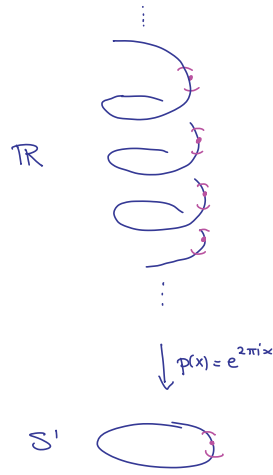


Figure 96: We can picture the exponential map as defining an infinite sheeted spiral cover over the circle.

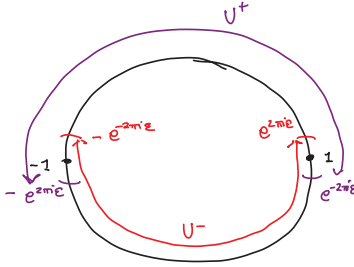


Figure 97: To prove the exponential map defines a fibration we choose a cover $U^{\pm}$ of the circle as shown.

Example: Let us return to our basic example: $E = \mathbb{R}$, $B = S^1$, and $p(x) = e^{2\pi i x}$. It can be pictured as a spiral covering the circle as in Figure 96. Suppose we choose $Y = \{y_0\}$ to be a single point. Then a map $\bar{f}$ amounts to choosing a single point on $\bar{f}(y_0) = z_0 \in S^1$, which

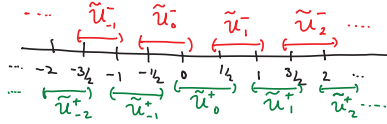


Figure 98: The inverse image of the open sets $U^\pm$ is a disjoint union of intervals, each interval is homeomorphic to U^+ or U^- under the exponential map.

we think of as a complex number with $|z_0| = 1$. The lift $f : Y \rightarrow E$ of this map is a choice of a real number $f(y_0) = x_0 \in \mathbb{R}$ so that $e^{2\pi i x_0} = z_0$. Now a homotopy $\bar{F} : I \times Y \rightarrow B = S^1$ is a path $z(t)$ in the unit circle with $z(0) = z_0$. The homotopy lifting property says that we can find $F : I \times Y \rightarrow \mathbb{R}$, which amounts to a function $F(t, y_0) = x(t)$ such that $x(0) = x_0$ and $\exp[2\pi i x(t)] = z(t)$. In terms of the figure 96, after choosing an initial point x_0 on the spiral above z_0 , as $z(t)$ moves continuously, the covering point $x(t)$ moves continuously up and down the spiral.

More generally, what we want to show is this: Suppose $\bar{f} : Y \rightarrow S^1$ is a continuous function such that we can also define a *continuous* function $f(y) = \theta(y) \in \mathbb{R}$ so that $\bar{f}(y) = e^{2\pi i \theta(y)}$. Now, suppose we have a homotopy $\bar{f}_t(y) = \bar{F}(t, y)$. The homotopy lifting property asserts that we can find a continuous real-valued function $F(t, y)$ so that if we define $\theta_t(y) := F(t, y)$ then we have $\bar{f}_t(y) = e^{2\pi i \theta_t(y)}$. We will now prove that this is indeed the case: ⁷⁴

The map p is an example of a “covering map,” defined precisely below. The most important consequence of this is that around any point in the circle there is an open neighborhood so that the inverse image under p is a disjoint union of neighborhoods in $\mathbb{R}$ each of which is homeomorphic to the neighborhood in S^1 .

To be very concrete choose U^+ to be the open arc through the upper half-plane from $e^{-2\pi i \epsilon}$ to $e^{2\pi i \epsilon}$ and let U^- be the arc through the lower half-plane from $e^{2\pi i \epsilon}$ to $e^{-2\pi i \epsilon}$. See Figure 97. We can take $0 < \epsilon < \frac{1}{4}$, but just think of ϵ as small and positive. Then the

⁷⁴Our proof is adapted from A. Hatcher, Algebraic Topology, Theorem 1.7.

inverse image of U^+ is a disjoint union of intervals, each of which is homeomorphic to U^+ :

$$p^{-1}(U^+) = \coprod_{n \in \mathbb{Z}} (x_- + n, x_+ + n) := \coprod_{n \in \mathbb{Z}} \tilde{U}_n^+ \quad (13.5)$$

where $x_+ = -\epsilon$ and $x_- = \frac{1}{2} + \epsilon$, while

$$p^{-1}(U^-) = \coprod_{n \in \mathbb{Z}} (y_- + n, y_+ + n) := \coprod_{n \in \mathbb{Z}} \tilde{U}_n^- \quad (13.6)$$

where $y_- = -\frac{1}{2} - \epsilon$ and $y_+ = \epsilon$. See Figure 98.

Now, for any $(t, y) \in [0, 1] \times Y$ there is a neighborhood $(t, y) \in (a(t, y), b(t, y)) \times \mathcal{O}(t, y)$ so that $\bar{F}$ maps this neighborhood entirely into U^+ or entirely into U^- .⁷⁵ Because $[0, 1]$ is compact we know that, for each $y \in Y$ there exists:

1. A neighborhood $\mathcal{O}(y)$, together with
2. A collection of points $0 = t_0 < t_1 < \dots < t_m = 1$

so that

$$\bar{F}([t_i, t_{i+1}] \times \mathcal{O}(y)) \subset U^{\alpha(y, i)} \quad (13.7)$$

where $U^{\alpha(y, i)}$ is either U^+ or U^- . When y is understood we will just denote $\alpha(y, i)$ by α_i .

Working at fixed y , let us consider first $i = 0$. Then the *lift* $f(y) = F(0, y)$ takes $\{0\} \times \mathcal{O}(y)$ into $\tilde{U}_{n_0}^{\alpha_0}$ for some definite integer n_0 . But now,

$$p : \tilde{U}_{n_0}^{\alpha_0} \rightarrow U^{\alpha_0} \quad (13.8)$$

is a homeomorphism. Denote its inverse by $q_{n_0}^{\alpha_0} : U^{\alpha_0} \rightarrow \tilde{U}_{n_0}^{\alpha_0}$. We can now define $F(t, y)$ on the set $[t_0, t_1] \times \mathcal{O}(y)$ by

$$F = q_{n_0}^{\alpha_0} \circ \bar{F} \quad (13.9)$$

In fact, since we require $p \circ F = \bar{F}$, and since p is locally inverted by $q_{n_0}^{\alpha_0}$ on $\tilde{U}_{n_0}^{\alpha_0}$, we *must* define F this way. Thus, the lift F on the set $[t_0, t_1] \times \mathcal{O}(y)$ is uniquely determined. We now try to extend this, for fixed y along the rest of the interval:

We know that $\bar{F} : [t_1, t_2] \times \mathcal{O}(y) \rightarrow U^{\alpha_1}$ for some $\alpha_1 \in \{\pm\}$. Now it might well happen that $\alpha_1 \neq \alpha_0$, but we do know that $\bar{F} : \{t_1\} \times \mathcal{O}(y) \rightarrow U^{\alpha_0} \cap U^{\alpha_1}$ and hence there is a *unique* integer n_1 so that

$$F(\{t_1\} \times \mathcal{O}(y)) \subset \tilde{U}_{n_0}^{\alpha_0} \cap \tilde{U}_{n_1}^{\alpha_1} \quad (13.10)$$

But now, $p : \tilde{U}_{n_1}^{\alpha_1} \rightarrow U^{\alpha_1}$ is a homeomorphism. Denote the inverse homeomorphism by $q_{n_1}^{\alpha_1} : U^{\alpha_1} \rightarrow \tilde{U}_{n_1}^{\alpha_1}$. Again, by continuity, we *must* define $F(t, y)$ on $[t_1, t_2] \times \mathcal{O}(y)$ by

$$F = q_{n_1}^{\alpha_1} \circ \bar{F} \quad (13.11)$$

We have now defined F on $[0, t_2] \times \mathcal{O}(y)$. We can continue the process in this way, defining a definite α_i, n_i and hence a definite local inverse to p , denoted $q_{n_i}^{\alpha_i}$ so that

$$F = q_{n_i}^{\alpha_i} \circ \bar{F} \quad (13.12)$$

⁷⁵Note that if $t = 0$ we should allow $a(t, y) < 0$ and interpret $(a(t, y), b(t, y))$ as $(a(t, y), b(t, y)) \cap [0, 1]$, and similarly for $t = 1$.

on $[t_i, t_{i+1}] \times \mathcal{O}(y)$. Since we have an open neighborhood $\mathcal{O}(y)$ for each $y \in Y$ we have in fact defined F on all of $[0, 1] \times Y$. The above construction made clear the necessity of the choices of lifting, and hence F is unique. ♠

Remarks

1. In general, for a fibration, the lift of a homotopy need not be unique. The above example showed that the lift is unique, but this is special to the case of coverings. See more below.
2. The fiber over a point $b \in B$ is the inverse image $p^{-1}(b)$, also denoted E_b . One can show that if B is path connected then any two fibers E_{b_1} and E_{b_2} are homotopy equivalent. The idea is to choose a path γ from b_1 to b_2 , and use the homotopy lifting property to define a map $U(\gamma) : E_{b_1} \rightarrow E_{b_2}$. Then homotopic paths lead to homotopic maps $U(\gamma_1) \sim U(\gamma_2)$ if $\gamma_1 \sim \gamma_2$. But this means that $U(\bar{\gamma})$ is a homotopy inverse to $U(\gamma)$. For more details see the proof of Proposition 16.3(a) in Bott and Tu, p. 200.

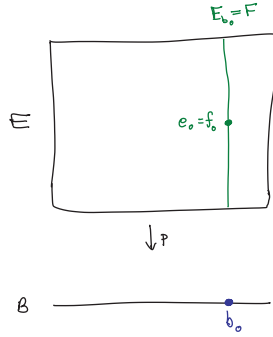


Figure 99: Choosing basepoints in a fibration. We use “ p ” for the projection to avoid confusion with homotopy groups, which are always denoted by π .

13.3 The long exact sequence of homotopy groups for a fibration

Let us now consider again a fibration $p : E \rightarrow B$. In this section we are going to describe a very interesting and beautiful relation between the homotopy groups of E and B . It can be very useful in computations.

To define homotopy groups we need to choose basepoints. We choose $b_0 \in B$, and $e_0 \in E$ so that $p(e_0) = b_0$. Let $\mathcal{F} = p^{-1}(b_0) := E_{b_0}$ be the fiber above b_0 . Let

$$\iota : \mathcal{F} \rightarrow E \tag{13.13}$$

be the inclusion, as in Figure 99. We denote the basepoint ν_0 of $\mathcal{F}$ to be e_0 so that $\iota(\nu_0) = e_0$.

Recall that $\pi_k(B, b_0)$ is the set of homotopy classes of maps $f : (I^k, \partial I^k) \rightarrow (B, b_0)$ and is a group for $k > 0$, while for $k = 0$, $\pi_0(B)$ is just the set of connected components. Since B has a distinguished basepoint there is a distinguished element of $\pi_0(B)$.

The inclusion and projection maps induce

$$\pi_k(\mathcal{F}, \nu_0) \xrightarrow{\iota_*} \pi_k(E, e_0) \xrightarrow{p_*} \pi_k(B, b_0) \quad (13.14)$$

These are group homomorphisms for $k > 0$.

Proposition The sequence (13.14) is *exact*, that is, $\ker p_* = \text{Im } \iota_*$.

Proof: First, since $p \circ \iota$ takes all of $\mathcal{F}$ to b_0 it is clear that $\text{Im } \iota_* \subset \ker p_*$. Now, suppose that $[f] \in \ker p_*$. Then f and $\bar{f} := p \circ f$ fit in the diagram:

$$\begin{array}{ccc} & (E, e_0) & \\ f \nearrow & \downarrow p & \\ (I^k, \partial I^k) & \xrightarrow{\bar{f}} & (B, b_0) \end{array} \quad (13.15)$$

Now $[f]$ being in the kernel of p_* simply means that $\bar{f}$ can be homotoped to the constant map. That is, there is a continuous family $\bar{f}_t$ with $\bar{f}_0 = \bar{f}$ while $\bar{f}_1(\vec{t}) = b_0$ is just the constant map. By the homotopy lifting property there is a lifting $f_t : (I^k, \partial I^k) \rightarrow (E, e_0)$. Since f_1 lifts the constant map, its image must lie in the fiber E_{b_0} . Therefore, $[f_1] \in \text{Im } \iota_*$.

♠

The surprising and important point is that there is a *connecting homomorphism*

$$\pi_k(B, b_0) \xrightarrow{\partial_k} \pi_{k-1}(\mathcal{F}, \nu_0) \quad (13.16)$$

giving a long exact sequence:

$$\begin{aligned} \cdots \rightarrow \pi_{k+1}(\mathcal{F}, \nu_0) \rightarrow \pi_{k+1}(E, e_0) \rightarrow \pi_{k+1}(B, b_0) \xrightarrow{\partial_{k+1}} \pi_k(\mathcal{F}, \nu_0) \xrightarrow{\iota_*} \pi_k(E, e_0) \xrightarrow{p_*} \pi_k(B, b_0) \xrightarrow{\partial_k} \pi_{k-1}(\mathcal{F}, \nu_0) \cdots \\ \cdots \rightarrow \pi_1(E, e_0) \xrightarrow{\partial_1} \pi_1(B, b_0) \rightarrow \pi_0(\mathcal{F}) \rightarrow \pi_0(E) \rightarrow \pi_0(B) \rightarrow 0. \end{aligned} \quad (13.17)$$

Note that the last three entries are in general just sets, and not groups, and there is no choice of basepoint.

Let us describe the construction of ∂_k . Suppose $g \in \pi_k(B, b_0)$. Then we choose a representative $g = [\bar{F}]$, where $\bar{F}$ is a map $\bar{F} : I^k \rightarrow B$ taking ∂I^k to b_0 . Decompose $I^k = I \times I^{k-1}$, with coordinates $(t, \vec{s})$ as in Figure 100. We view $\bar{F}$ as a homotopy of maps from $(I^{k-1}, \partial I^{k-1}) \rightarrow (B, b_0)$. That is, at fixed time t , we define $\bar{f}_t(\vec{s}) := \bar{F}(t, \vec{s})$, so that $\bar{f}_t$ is a map $(I^{k-1}, \partial I^{k-1}) \rightarrow (B, b_0)$. Recall that ∂I^{k-1} is the boundary where at least one

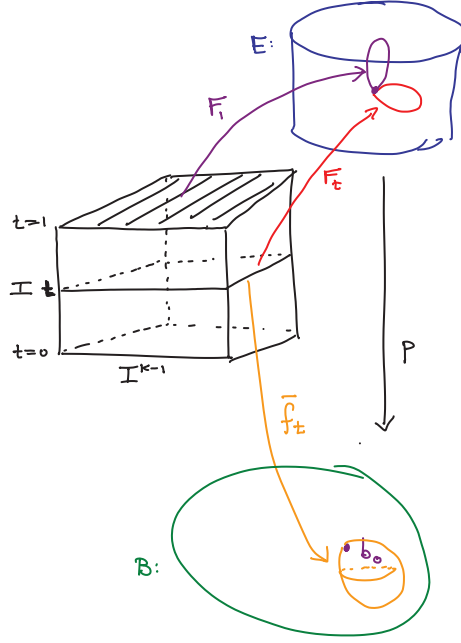


Figure 100: Construction of the connecting homomorphism. We view a map $(I^k, \partial I^k) \rightarrow (B, b_0)$ as a family of maps $(I^{k-1}, \partial I^{k-1}) \rightarrow (B, b_0)$ by considering the first coordinate to be time. We now attempt to lift the homotopy. We choose the bottom face of the box to map to e_0 . Then, using the homotopy lifting property we can lift the map from the box into the total space (E, e_0) . The top face of the box must cover the constant map, but need not be the constant map. It defines a map of I^{k-1} into the fiber over b_0 .

component of $\vec{s}$ is zero or one. For each t , $\bar{f}_t$ and maps to b_0 . Note too that $\bar{f}_0(\vec{s}) = b_0$ and $\bar{f}_1(\vec{s}) = b_0$ are the constant map to b_0 . Now, since $\bar{f}_0$ is the constant map it is easy to find a lifting: we simply choose the constant map $f_0 : I^{k-1} \rightarrow E$ given by the constant e_0 . Now we invoke the homotopy lifting property with $Y = I^{k-1}/\partial I^{k-1} = S^{k-1}$, where $k \geq 1$ and S^{k-1} has a distinguished point y_0 given by collapsing ∂I^{k-1} . The homotopy $\bar{F}$ has a lift given by a continuous map $F : I \times S^{k-1} \rightarrow E$ such that $F(0, \vec{s}) = e_0$, is the lift of $\bar{F}(0, \vec{s})$. However, the homotopy lifting property does *not* say that at $t = 1$ the lift should be constant! Indeed, $f_1(\vec{s}) := F(1, \vec{s})$ is a map $(S^{k-1}, pt) \rightarrow (E, e_0)$ or, equivalently,

$$f_1 : (I^{k-1}, \partial I^{k-1}) \rightarrow (E, e_0) \quad (13.18)$$

Now, $p \circ F = \bar{F}$, so $p \circ f_1 = \bar{f}_1$ and $\bar{f}_1$ is the constant map to b_0 . Therefore, f_1 must map all of I^{k-1} into the fiber $\mathcal{F} = E_{b_0}$. Therefore f_1 defines a homotopy class

$$[f_1] \in \pi_{k-1}(\mathcal{F}, \nu_0) \quad (13.19)$$

We define the connecting homomorphism to be:

$$\partial_k([\bar{F}]) = [f_1] \quad (13.20)$$

Of course, for this to make sense one must show several *choices* which were made in the above construction do not affect the definition of ∂_k , if it is to be well-defined. The choices are:

1. A choice of representative $\bar{F}$ of a class in $\pi_k(B, b_0)$.
2. A choice of homotopy lifting F of $\bar{F}$.

We leave to the reader the demonstration that ∂_k is indeed well-defined.

The connecting homomorphism is subtle and nontrivial. Another useful picture explaining what it does is the following. We identify $I^k/\partial I^k$ with S^k/pt .

1. For $k = 1$ the homotopy $\bar{f}_t$ is a family of maps from a point into (B, b_0) with $\bar{f}_0 = \bar{f}_1 = b_0$. That is, it is a closed path. The lift is a family of maps f_t into E with $f_0 = e_0$, but, crucially, f_1 need not be e_0 . It is just in the fiber E_{b_0} above b_0 . So, although the homotopy $\bar{F} : (I, \partial I) \rightarrow (B, b_0)$ defines a closed path in B , the lifted homotopy $F : I \rightarrow E$ has $F(0) = e_0$, but $F(1) \in E_{b_0}$ need not be the same as e_0 , and hence the lifted path need not be closed: $\partial_1[\bar{F}] = [F(1)] \in \pi_0(\mathcal{F})$. This phenomenon is known as *monodromy*.
2. For $k = 2$, $\bar{f}$ maps a square into B with the boundaries of the square mapping into b_0 . The quotient $I^2/\partial I^2$ is an S^2 with a distinguished point, and we can identify the “constant time” slices as a lasso opening up and sliding over the sphere as in Figure 101. At fixed t , $\bar{f}_t$ is a loop in B based at b_0 . The lift F_t of $\bar{f}_t$ is similarly a loop in E based at e_0 . Although the lift F_0 of $\bar{f}_0$ is just the constant loop at e_0 , the loop F_1 need not be the constant loop. It just sits in the fiber above b_0 . Then $\partial_2[\bar{f}] = [F_1] \in \pi_1(\mathcal{F}, \nu_0)$.
3. This picture generalizes. We consider a map $\bar{f} : (S^k, *) \rightarrow (B, b_0)$ to be a family of maps $\bar{f}_t : (S^{k-1}, *) \rightarrow (B, b_0)$. We view the lower dimensional spheres as a “lasso” around the sphere of one higher dimension, analogous to Figure 101.

We leave it as an exercise to prove exactness of (13.17) at the stages $\pi_k(\mathcal{F})$ and $\pi_k(B)$ in complete detail. Here are some hints:

1. If $f : (I^k, \partial I^k) \rightarrow (E, e_0)$ then $\bar{f} = p \circ f$ comes together with its lifting. So if we want to compute $\partial_k(p_*(f)) = \partial_k[\bar{f}]$ then since f is constant on all the boundaries of ∂I^k $\partial_k([\bar{f}])$ is represented by the constant map, and is clearly the trivial homotopy class. Therefore $\text{Im } p_* \subset \ker \partial_k$. Similarly, if $[\bar{f}]$ is in the kernel of ∂_k then the lifted homotopy $F_{t=1}$ can be further homotoped to a constant. Attaching this extra box to the one we already have gives a map of the box $(I^k, \partial I^k)$ into (E, e_0) covering the one into (B, b_0) , so $\ker \partial_k \subset \text{Im } p_*$. We conclude that $\ker \partial_k = \text{Im } p_*$, and the sequence is exact at $\pi_k(B, b_0)$.
2. Now, suppose $[f] \in \pi_k(\mathcal{F}, \nu_0)$ is in the image of ∂_{k+1} . That means there is a family of maps $F_t : (I^k, \partial I^k) \rightarrow (E, e_0)$ so that $F_0(\vec{s}) = e_0$ and $F_1(\vec{s}) = f(\vec{s})$. If we run F_t

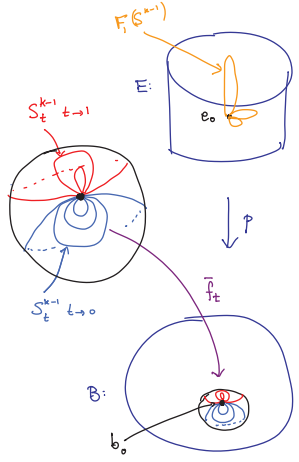


Figure 101: The connecting homomorphism for the case $\partial_2 : \pi_2(B, b_0) \rightarrow \pi_1(\mathcal{F}, \nu_0)$. The family of circles sweeps out a 2-sphere in B . For fixed t the circle in the base $\bar{\gamma}_t(s)$, $0 \leq s \leq 1$ lifts to a circle in the total space, $\gamma_t(s)$. As $t \rightarrow 1$ the circle in the base shrinks back down to a point, but the circle in the total space does not shrink. Rather, it wraps around the fiber above b_0 . This picture generalizes: A map of S^k into B can be considered to be a loop (homotopy) of maps S^{k-1} into B beginning and ending with the constant loop into b_0 . We lift the initial point loop to e_0 and then invoke the covering homotopy property to get a family of based maps S^{k-1} into the total space E . At the end of the homotopy, since the map downstairs goes to a point, the map upstairs must go into the fiber over that point .

backwards we obtain a homotopy of $\iota \circ f$ to the constant map. Note that this only applies to $\iota_*([f])$ and not to $[f]$ because F_t takes values in E . Thus $\text{Im } \partial_{k+1} \subset \ker \iota_*$. Now suppose $[f] \in \ker \iota_*$. Then there is a continuous family of maps $F_t : (I^k, \partial I^k) \rightarrow (E, e_0)$ so that $F_0 = f$ and $F_{t=1} = e_0$ is the constant map. Consider this as a single map $F : I \times I^k \rightarrow E$ and then $p \circ F$ takes ∂I^{k+1} to a single point b_0 and hence defines a homotopy class in $\pi_{k+1}(B, b_0)$. By construction, $[f]$ is ∂_{k+1} applied to this homotopy class. Hence $\ker \iota_* \subset \text{Im } \partial_{k+1}$. We conclude that $\ker \iota_* = \text{Im } \partial_{k+1}$ and the sequence is exact at $\pi_k(\mathcal{F}, \nu_0)$.

Example 1: Let us return to our fibration with $E = \mathbb{R}$ and $B = S^1 = \mathbb{R}/\mathbb{Z}$. Since $\mathbb{R}$ is contractible and $\mathbb{Z}$ is discrete the exact homotopy sequence tells us that

$$\pi_k(S^1, b_0) = \delta_{k,1} \mathbb{Z} \quad (13.21)$$

For $k = 1$ consider $f_n : (I, \partial I) \rightarrow (S^1, 1)$ given by $f_n(t) = \exp[2\pi i n t]$ with $n \in \mathbb{Z}$. It defines a homotopy class $[f_n] \in \pi_1(S^1, 1)$. Take $b_0 = 1$, choose a lift of $f_n(0)$ in $\mathbb{R}$, call it $e_0 \in \mathbb{Z}$. Then the lifted homotopy is clearly $F_t = e_0 + nt$ at $t = 1$ this is the map $F_1 = e_0 + n$. Therefore $\partial_1[f_n]$ is in the component labeled by n , if we take $\nu_0 = e_0$ to be the basepoint in $F = \mathbb{Z}$.

Example 2. As we will prove later, $SU(n)$ is a topological group, and it is connected and simply connected. Let us consider the quotient group $SU(n)/\mathcal{Z}$ where $\mathcal{Z}$ is the subgroup given by the *center* of $SU(n)$:

$$\mathcal{Z} = \{\omega^j \mathbf{1}_{n \times n} | j = 0, \dots, n-1\} \quad (13.22)$$

where ω is a primitive n^{th} root of 1. This subgroup is isomorphic to $\mathbb{Z}_n$. The projection $p : SU(n) \rightarrow SU(n)/\mathcal{Z}$ is a fibration. In fact, it is a covering map (see below). By the homotopy exact sequence we have

$$\cdots \rightarrow \pi_1(SU(n)) \rightarrow \pi_1(SU(n)/\mathcal{Z}) \rightarrow \pi_0(\mathbb{Z}_n) \rightarrow \cdots \quad (13.23)$$

and hence we conclude

$$\pi_1(SU(n)/\mathcal{Z}) \cong \mathbb{Z}_n. \quad (13.24)$$

An example of a nontrivial loop in $SU(n)/\mathcal{Z}$ is

$$g(\phi) = [Diag\{e^{i\phi/n}, e^{i\phi/n}, \dots, e^{-i\phi \frac{n-1}{n}}\}] \quad (13.25)$$

where the equivalence class $[\cdot]$ denotes the projection to $SU(n)/\mathcal{Z}$. Note that $g(\phi + 2\pi) = g(\phi)$ in $SU(n)/\mathcal{Z}$, but that this is *not* a closed loop in $SU(n)$. If we try to lift to

$$\tilde{g}(\phi) = Diag\{e^{i\phi/n}, e^{i\phi/n}, \dots, e^{-i\phi \frac{n-1}{n}}\} \quad (13.26)$$

then $\tilde{g}(\phi + 2\pi) = \tilde{g}(\phi)\omega$, and indeed $\partial_1[g(\phi)] = \omega \mathbf{1}_{n \times n}$.

Example 3 *The Hopf fibration.* We will prove later that there is a fibration (in fact, a fiber bundle) $p : S^3 \rightarrow S^2$.⁷⁶ Let us just accept the existence of this map for the moment and see what the sequence (13.17) tells us.

$$\begin{aligned} \rightarrow \pi_3(S^1) &\rightarrow \pi_3(S^3) \rightarrow \pi_3(S^2) \rightarrow \\ \rightarrow \pi_2(S^1) &\rightarrow \pi_2(S^3) \rightarrow \pi_2(S^2) \rightarrow \\ \rightarrow \pi_1(S^1) &\rightarrow \pi_1(S^3) \rightarrow \pi_1(S^2) \rightarrow \end{aligned} \quad (13.27)$$

Now let us use our knowledge that $\pi_k(S^1, 1) = \delta_{k,1}\mathbb{Z}$, as well as the homotopy groups $\pi_k(S^n) = \delta_{k,n}\mathbb{Z}$ for $0 < k \leq n$ to simplify this to

$$\begin{aligned} \rightarrow 0 &\rightarrow \mathbb{Z} \rightarrow \pi_3(S^2) \rightarrow \\ \rightarrow 0 &\rightarrow 0 \rightarrow \mathbb{Z} \rightarrow \\ \rightarrow \mathbb{Z} &\rightarrow 0 \rightarrow 0 \rightarrow \end{aligned} \quad (13.28)$$

⁷⁶Here is a lightning summary: Observe that the most general 2×2 traceless hermitian matrix is of the form $\vec{x} \cdot \vec{\sigma}$, where $\vec{x} \in \mathbb{R}^3$. Moreover, $-\det(\vec{x} \cdot \vec{\sigma}) = \vec{x}^2$ is the Euclidean norm. Now, define a homomorphism from $\rho : SU(2) \rightarrow SO(3)$ by declaring $\rho(u)$ to be the rotation $\vec{x} \rightarrow \vec{x}'$ such that $u\vec{x} \cdot \vec{\sigma}u^{-1} = \vec{x}' \cdot \vec{\sigma}$. Note that $u\vec{x} \cdot \vec{\sigma}u^{-1}$ is traceless hermitian, so $\vec{x}'$ exists, and has the same determinant, so $\vec{x}^2 = (\vec{x}')^2$. Now take the rotation $\rho(u)$ and rotate the north pole on S^2 with this matrix. The fiber of this map can be seen to be a circle double covering the circle of rotations around the axis through the north pole. Finally, we note that $SU(2)$, as a manifold, is precisely S^3 .

By exactness of the sequence we learn that:

$$\pi_3(S^2) = \mathbb{Z} \tag{13.29}$$

Notice the striking contrast with homology theory, where $H_3(S^2) = 0$. This discovery came as quite a shock when Hopf discovered it in the 1930's!

Remarks:

1. The fibers E_b of a fibration for different values of b need not be homeomorphic, but they do have the same homotopy type, if B is path connected. We prove this in Section **** below. For the moment note that it would be very strange if this were not true since we would potentially have different groups $\pi_k(\mathcal{F})$ in the LES with the same $\pi_k(B)$ and $\pi_k(E)$.
2. The last three maps in (13.16) are not group homomorphisms since in general $\pi_0(X)$ has no natural group structure. However, since we have chosen basepoints they are pointed sets. We can continue to understand the sequence as being exact if we interpret the kernel of a map between pointed sets to be the inverse image of the basepoint. That is, the basepoints ν_0, e_0, b_0 of $\mathcal{F}, E, B$ define a particular component, hence a particular element of $\pi_0(\mathcal{F}), \pi_0(E), \pi_0(B)$ which serves as the identity. Thus exactness says:
 1. Every component of B is covered by some component of E .
 2. The components of E covering that component of B containing the basepoint b_0 are all connected components of $\mathcal{F} = E_{b_0}$ (this is obvious),
 3. The components of $\mathcal{F} = E_{b_0}$ which become connected to e_0 , when embedded in E are precisely the monodromy lifts of elements of $\pi_1(B, b_0)$. Again, this is clear: The monodromy lifts provide the connecting paths.
3. Although π_0 in general is not a group, when F, E, B are all groups then since $\pi_0(G)$ is naturally a group it makes sense to interpret the last three maps of (13.16) as group homomorphisms.
4. *The path fibration* One of the very beautiful application of (13.17) is to the path fibration $p : (\mathcal{P}(X, x_0), \alpha_0) \rightarrow (X, x_0)$ where $\alpha_0(t) = x_0$ is the constant path. The map p is defined by taking the endpoint of the path: $p(\alpha) = \alpha(1)$. This can be shown to be a fibration. As we have seen, $\mathcal{P}(X, x_0)$ is contractible to the point α_0 . But, the fiber has homotopy type of $\Omega_{x_0}(X)$, the based loops in X . We therefore conclude that

$$\pi_k(\Omega_{x_0}(X)) = \pi_{k+1}(X, x_0) \tag{13.30}$$

5. *The Hopf invariant.* For any map $f : S^3 \rightarrow S^2$ we can associate an integer topological invariant since f induces a homomorphism $f_* : \pi_3(S^3) \rightarrow \pi_3(S^2)$, and hence a homomorphism $\mathbb{Z} \rightarrow \mathbb{Z}$, and $\text{Hom}(\mathbb{Z}, \mathbb{Z}) \cong \mathbb{Z}$. This is called the Hopf invariant $H(f)$ and it turns out that it can be written as an integration of a local density, in a way analogous to the way the degree of a map $g : S^n \rightarrow S^n$ can be. Suppose $f : S^3 \rightarrow S^2$. Let ω be 2-form on S^2 representing the generator of $H^2(S^2, \mathbb{Z})$. For example, we could let ω be the unit volume form on S^2 . Now, $H^2(S^3, \mathbb{Z}) = 0$. Therefore, $f^*(\omega) = d\Theta$ for some globally well-defined 1-form Θ on S^3 . Then we define:

$$H(f) = \int_{S^3} \Theta \wedge d\Theta \quad (13.31)$$

In the case when f is a projection in a principal $U(1)$ fibration we will interpret Θ below as a $U(1)$ connection and the Hopf invariant as a Chern-Simons invariant. We will see that for $U(1)$ bundles corresponding to $n \in \pi_1(S^1)$ the Hopf invariant is just n .

Some sources:

1. Bredon, Corollary 6.12
2. Bott and Tu, section 16
3. Husemoller, ch. 1

Exercise *Choices, choices*

Check that the connecting homomorphism is well-defined $\partial_k[\bar{F}] = [f_1]$.

Exercise

Show that $\pi_k(S^3) = \pi_k(S^2)$ for $k \geq 3$.

Exercise

Consider the circle as the unit circle in the complex plane. Consider the map $p : S^1 \rightarrow S^1$ given by $p(z) = z^n$, with n an integer.

- a.) Prove this is a fibration.
 - b.) Work through the LES sequence in homotopy for this fibration.
-

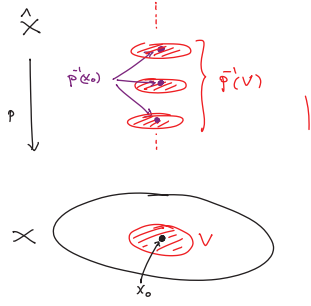


Figure 102: The pancake picture of a covering map

13.4 Covering spaces

A very important special class of fibrations are the *covering spaces*. These are simply the fibrations $p : E \rightarrow B$ such that the fiber is a discrete set. The precise definition is the following:

Definition. A continuous map $p : \hat{X} \rightarrow X$ of topological spaces is a covering map if the inverse images under p of sufficiently small neighborhoods $V \subset X$ is a disjoint union of homeomorphic copies of V .

We can write state the covering map condition as the condition that for every $x \in X$ there is a neighborhood V of x such that

$$p^{-1}(V) = \coprod_{\alpha \in \mathcal{F}} \tilde{V}_\alpha \quad (13.32)$$

where α runs over the fiber $\mathcal{F}$ above x . $\mathcal{F}$ is a set with discrete topology and, for each α , the map $p : \tilde{V}_\alpha \rightarrow V$ is a homeomorphism. See figure 102.

We claim that a covering map has the homotopy lifting property and therefore a covering is a special case of a fibration. The proof is completely parallel to the proof we used before that $p : \mathbb{R} \rightarrow S^1$, given by the exponential map, is a fibration: We cover X by open sets U^α so that for each α the inverse image $p^{-1}(U^\alpha)$ is a disjoint union of homeomorphic copies:

$$p^{-1}(U^\alpha) = \coprod_\nu \tilde{U}_\nu^\alpha \quad (13.33)$$

where now the ν run over some index set, not necessarily the integers. Given the data of a continuous map $\bar{F} : I \times Y \rightarrow X$ and a lift $f : Y \rightarrow \hat{X}$ of $\bar{f}(y) := \bar{F}(0, y)$ we proceed as before: For each $(t, y) \in I \times Y$ we find neighborhoods $(a(t, y), b(t, y)) \times \mathcal{O}(t, y)$ so that

$\bar{F}((a(t, y), b(t, y)) \times \mathcal{O}(t, y)) \subset U^\alpha$ for some α . Again, using compactness of I we have a neighborhood $\mathcal{O}(y)$ of $y \in Y$ and times (which might depend on y) $0 = t_0 < t_1 < \dots < t_m = 1$ so that

$$\bar{F}([t_i, t_{i+1}] \times \mathcal{O}(y)) \subset U^{\alpha(y, i)} \quad (13.34)$$

Then we start at $i = 0$ and use the lift f to distinguish a cover $\tilde{U}_{\nu_0}^{\alpha(y, 0)}$ into which it maps so that we are forced to define

$$F := q_{\nu_0}^{\alpha(y, 0)} \circ \bar{F} \quad (13.35)$$

on $[0, t_1] \times \mathcal{O}(y)$. Next $\bar{F} : [t_1, t_2] \times \mathcal{O}(y) \rightarrow U^{\alpha(y, 1)}$, and therefore $\bar{F}(\{t_1\} \times \mathcal{O}(y)) \subset U^{\alpha(y, 0)} \cap U^{\alpha(y, 1)}$. Therefore, the lift must satisfy

$$F(\{t_1\} \times \mathcal{O}(y)) \subset \tilde{U}_{\nu_0}^{\alpha(y, 0)} \cap \tilde{U}_{\nu_1}^{\alpha(y, 1)} \quad (13.36)$$

for some ν_1 . But again p has a local inverse $q_{\nu_1}^{\alpha(y, 1)} : U^{\alpha(y, 1)} \rightarrow \tilde{U}_{\nu_1}^{\alpha(y, 1)}$ and hence we extend F to $[t_1, t_2] \times \mathcal{O}(y)$ by

$$F = q_{\nu_1}^{\alpha(y, 1)} \circ \bar{F} \quad (13.37)$$

We can continue this way to define $F : [0, 1] \times \mathcal{O}(y) \rightarrow \hat{X}$ and by uniqueness, the construction is well defined on $\mathcal{O}(y) \cap \mathcal{O}(y')$.

For a covering space the long exact sequence in homotopy collapses since $\pi_j(\mathcal{F}) = 0$ for $j > 0$ while $\pi_0(\mathcal{F})$ is just the set of points in $\mathcal{F}$, or equivalently, the set of connected components of $\mathcal{F}$.

Applying the homotopy exact sequence we conclude immediately that

$$\pi_k(\hat{X}, \hat{x}_0) = \pi_k(X, x_0) \quad k > 1, \quad (13.38)$$

and moreover if X is connected:

$$1 \rightarrow \pi_1(\hat{X}, \hat{x}_0) \xrightarrow{p_*} \pi_1(X, x_0) \xrightarrow{\partial} \pi_0(\mathcal{F}) \rightarrow \pi_0(\hat{X}) \rightarrow 1 \quad (13.39)$$

From this we learn that $p_* : \pi_1(\hat{X}, \hat{x}_0) \rightarrow \pi_1(X, x_0)$ is an injection.

Moreover, if $\hat{X}$ is connected then the index of $p_*(\pi_1(\hat{X}, \hat{x}_0))$ in $\pi_1(X, x_0)$ is the number of sheets of the cover. Since $\pi_0(\mathcal{F})$ is not a group, in general, the last arrow simply means that the map ∂ is surjective. After all, if X is connected then each component of E must contain some inverse image of x_0 . Moreover, exactness at $\pi_0(\mathcal{F})$ means that the inverse images in $p^{-1}(x_0)$ in a fixed connected component of E are all in the image of ∂ .

Exercise

Show that the quotient map $p : S^2 \rightarrow \mathbb{RP}^2$ is a covering map and use the LES to compute $\pi_1(\mathbb{RP}^2)$.

Exercise

Show that the map $p : \mathbb{C}^* \rightarrow \mathbb{C}^*$ given by the holomorphic function $p(z) = z^n$ (with n any nonzero integer) is a covering map. Describe how the LES works for this map.

13.5 Path lifting, connections, monodromy, and differential equations

When we proved the homotopy lifting property for coverings we actually proved something stronger than what the definition of a fibration demands: The lifting is *unique*. Let us apply this to paths in X :

We can consider a path $\varphi : [0, 1] \rightarrow X$ with $\varphi(0) = x_0$ to be a homotopy of the constant map. That is, we choose Y to be a single point: $Y = \{y_0\}$ and $\bar{f}(y_0) = x_0$. Then a lift $f : Y \rightarrow \hat{X}$ of $\bar{f}$ is simply a choice of a point $\hat{x}_0$ in the fiber above x_0 . Thus,

Given a path $\varphi(t)$ in X beginning at x_0 and a choice of preimage $\hat{x}_0$ there is a unique lifted path $\hat{\varphi}(t)$ in $\hat{X}$ which lifts $\varphi(t)$, that is, such that $p(\hat{\varphi}(t)) = \varphi(t)$.

Moreover, this lifting property is nicely compatible with our composition of paths. Suppose φ_1 is a continuous path from x_0 to x_1 and φ_2 is a continuous path from x_1 to x_2 . Choose a lift $\hat{x}_0$ of x_0 . Then there is a unique lift $\hat{\varphi}_1$ of φ_1 with initial point $\hat{x}_0$. Now, the endpoint $\hat{x}_1 := \hat{\varphi}_1(1)$ provides a lift of $\varphi_1(1) = \varphi_2(0) = x_1$. Using this lift we have a unique lifting $\hat{\varphi}_2$ of φ_2 , and it should be clear that:

$\hat{\varphi}_1 \star \hat{\varphi}_2$ is the unique lift of $\varphi_1 \star \varphi_2$ with initial point $\hat{x}_0$.

In general, we have the following definition:

Definition: A *connection* on a fibration $p : E \rightarrow B$ is a rule for lifting paths on B to paths on E which is consistent with composition of paths. That is: Given a path $\varphi : [0, 1] \rightarrow B$ connecting b_0 to b_1 , and a preimage $e_0 \in p^{-1}(b_0)$, a connection assigns a unique path $\mathbb{P}(\varphi) : [0, 1] \rightarrow E$ such that

1. $\mathbb{P}(\varphi)(0) = e_0$.
2. $p(\mathbb{P}(\varphi)(t)) = \varphi(t)$.
3. $\mathbb{P}(\varphi_1 \star \varphi_2) = \mathbb{P}(\varphi_1) \star \mathbb{P}(\varphi_2)$ where the initial point of $\mathbb{P}(\varphi_2)$ is the endpoint of $\mathbb{P}(\varphi_1)$.

The lifted path $\mathbb{P}(\varphi)$ is sometimes called the *parallel transport*. What we have proven above is that for a covering space there is a unique connection compatible with p , namely $\mathbb{P}(\varphi) = \hat{\varphi}$. For more general fibrations there can be (infinitely) many connections.

In general, if φ is a *closed* path from b_0 to b_0 in B the lifted path $\mathbb{P}(\varphi)$ is *not* a closed path. That is

$$\mathbb{P}(\varphi)(1) \neq e_0 \tag{13.40}$$

When this happens the connection is said to have *monodromy around the path φ* and the map of the fiber $e_0 \rightarrow \mathbb{P}(\varphi)(1)$ is the monodromy map of the connection associated with the path φ .

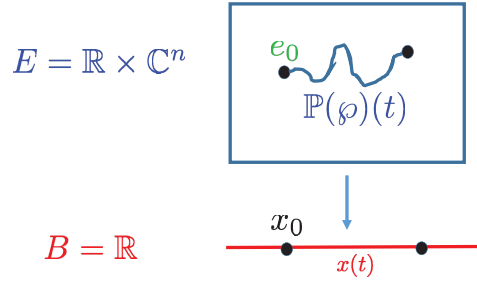


Figure 103: Lifting a path on $\mathbb{R}$ for the trivial bundle.

13.5.1 Interlude: Ordinary differential equations

People often associate gauge fields with connections. Indeed, the gauge fields of physics provide local data that give path lifting rules through differential equations.

Example 1: The most elementary example is obtained by taking $B = \mathbb{R}$ and $E = \mathbb{R} \times \mathbb{C}^n$, with $p : E \rightarrow B$ simply being projection onto the first factor, i.e. $p(x, \vec{v}) = x$. Now, given a path φ in B given by $x(t)$, $0 \leq t \leq 1$ we can define a path-lifting rule, i.e., a connection, by choosing a function of x valued in $n \times n$ complex matrices. Let us call it $A(x)$. Then the path-lifting rule is

1. If φ is a path from $\varphi : x_0 \rightsquigarrow x_1$, choose an element of the fiber $e_0 = (x_0, \vec{v}_0)$ above the initial point.
2. Then, solve the ordinary differential equation, with boundary condition provided by the lift e_0 of the initial point:

$$\frac{d}{dt} \vec{v}(t) = A(x(t)) \frac{dx}{dt} \vec{v}(t) \quad \vec{v}(0) = \vec{v}_0. \quad (13.41)$$

The rule for the lifted path (determined by the choice $A(x)$) is then

$$\mathbb{P}(\varphi)(t) := (x(t), \vec{v}(t)) \quad (13.42)$$

See Figure 103.

The reason that this rule is indeed compatible with composition of paths is that the equation (13.41) is invariant under reparametrization $t \rightarrow f(t)$ of the time t , so long as $f(t)$ is differentiable and $f'(t) > 0$. Using this fact and the existence and uniqueness of solutions to first order linear ODE's we have

$$\mathbb{P}(\varphi_1 \star \varphi_2) = \mathbb{P}(\varphi_1) \star \mathbb{P}(\varphi_2) \quad (13.43)$$

Remarks

1. We assume here that $\wp$ is a piecewise-differentiable path, i.e. $x(t)$ is a continuous function which is differentiable on intervals. (But the derivative can be discontinuous at isolated points.)
2. We assume that $A(x)$ is nonsingular on the path.
3. Note that since the differential equation is invariant under complex conjugation, if $A(x)$ is real then if $\vec{v}_0$ is real the solution $\vec{v}(t)$ will be real.

The path ordered exponential

As a matter of fact, one can write an “explicit” solution of the differential equation (13.41) and this representation can be useful. The solution is can be written as follows. Define an $n \times n$ complex matrix:

$$\mathbb{U}(\wp_t) := 1 + \int_0^t A(x(t_1))\dot{x}(t_1)dt_1 + \sum_{m=2}^{\infty} \int_0^t dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{m-1}} dt_m A(x(t_1))\dot{x}_1 \cdots A(x(t_m))\dot{x}_m \quad (13.44)$$

then we have

$$\vec{v}(t) = \mathbb{U}(\wp_t)\vec{v}_0. \quad (13.45)$$

To prove this, note that by explicit differentiation

$$\frac{d}{dt}\mathbb{U}(\wp_t) = A(x(t))\dot{x}(t)\mathbb{U}(\wp_t) \quad (13.46)$$

and note that $\mathbb{U}(\wp_0) = 1$.

Remarks

1. $\mathbb{U}(\wp_t)$ is an operator, independent of the choice of lift $\vec{v}_0$ of the initial point.
2. Matrix multiplication of $\mathbb{U}(\wp)$ is contravariant with respect to composition of paths:

$$\mathbb{U}(\wp_1 \star \wp_2) = \mathbb{U}(\wp_2)\mathbb{U}(\wp_1) \quad (13.47)$$

3. For the piecewise continuous path $\wp_t \star \bar{\wp}_t$ it is clear that the parallel transport takes $(x_0, \vec{v}_0) \rightarrow (x_0, \vec{v}_0)$. Therefore $\mathbb{U}(\wp_t \star \bar{\wp}_t) = 1_{n \times n}$. It follows that $\mathbb{U}(\wp_t)$ is an *invertible* matrix.
4. $\mathbb{U}(\wp_t)$ is invariant under reparametrizations of the path $x(t)$ by $t \rightarrow f(t)$ where $f'(t) > 0$. It therefore makes sense to write

$$\mathbb{U}(\wp_t) = 1 + \int_{x_0}^{x(t)} A(x_1)dx_1 + \sum_{m=2}^{\infty} \int_{x_0}^{x(t)} A(x_1)dx_1 \int_{x_0}^{x_1} dx_2 \cdots \int_{x_0}^{x_{m-1}} A(x_m)dx_m \quad (13.48)$$

This expression has a further useful representation by introducing the *(left) time ordered product* of matrices defined by

$$T_\ell(A(x(t_1)), \dots, A(x(t_m))) := A(x(t_{\sigma(1)})) \cdots A(x(t_{\sigma(m)})) \quad (13.49)$$

where $\sigma \in S_m$ is a permutation such that

$$t_{\sigma(1)} \geq \cdots \geq t_{\sigma(m)} \quad (13.50)$$

Note that if all the times are distinct then the permutation is unique. If some times coincide then σ is not uniquely determined, but any two permutations lead to the same RHS for (13.49). Using this notation we can write:

$$\mathbb{U}(\wp_t) = 1 + \sum_{m=1}^{\infty} \frac{1}{m!} \int_0^t dt_1 \dot{x}_1 \cdots \int_0^t dt_m \dot{x}_m T_\ell[A(x(t_1)), \dots, A(x(t_m))] \quad (13.51)$$

which motivates the notation

$$\mathbb{U}(\wp_t) := \text{Pexp} \int_0^t dt_1 \dot{x}_1 A(x(t_1)) \quad (13.52)$$

5. **Warning:** Do not confuse the path-ordered exponential with the ordinary exponential:

$$\exp \int_0^t dt_1 \dot{x}_1 A(x(t_1)) \quad (13.53)$$

If $A(x)$ is a family of *commuting* matrices then the exponential and the path-ordered exponential will be the same. In general they are different.

6. If one reversed the order of all the inequalities in (13.50) then (13.49) would define the (*right*) *time-ordered product* and the multiplication rule would be covariant. This is the matrix we would apply to the differential equation

$$\frac{d}{dt} \vec{v}(t) = \vec{v}(t) A(x(t)) \frac{dx}{dt} \quad \vec{v}(0) = \vec{v}_0. \quad (13.54)$$

where $\vec{v}(t)$ is a row vector.

Now we are not going to see interesting monodromy around closed paths in the above example, because $\pi_1(\mathbb{R}, x_0) = 0$, but a small modification of the above example produces interesting examples.

Example 2: We now take $B = S^1$ which we regard as both the unit circle in the complex plane and the quotient $\mathbb{R}/\mathbb{Z}$ given by identifying $x \sim x + 1$. Now we take

$$E = S^1 \times \mathbb{C}^n \quad (13.55)$$

and the projection p is simply projection onto the first factor: $p(z, \vec{v}) = z$. Now let $A(x)$ be a matrix-valued function as before but now impose the condition that it be periodic: $A(x+1) = A(x)$. Consider a path $\wp$ on S^1 with $\wp(0) = z_0$. The fiber above z_0 is the set of points $(z_0, \vec{v})$ where $\vec{v} \in \mathbb{C}^n$. As always, to lift the path $\wp$ to a path $\mathbb{P}(\wp)$ in E we must choose a lift $e_0 = (z_0, \vec{v}_0)$ of the initial point of the path. Now, to write the differential equation we also lift the path $z(t)$ by choosing some initial point x_0 so that $z(t) = \exp[2\pi i x(t)]$, with $x(0) = x_0$. Of course, if $\wp$ is a closed path so that $z(1) = z_0$ then

$x(t)$ need not be closed, but rather $x(1) = x_0 + n$, where n is the winding number of φ . Now we consider exactly the same differential equation (13.41), and we produce a family $(x(t), \vec{v}(t))$. It now makes sense to pass from $x(t)$ to $z(t)$ precisely because $A(x)$ is periodic in x , so now the lifted path is:

$$\mathbb{P}(\varphi)(t) = (z(t), \vec{v}(t)) \quad (13.56)$$

Even though $z(1) = z(0) = z_0$ and $A(x(1)) = A(x_0 + n) = A(x(0))$ are single-valued, there is no reason for $\vec{v}(t)$ to be single valued. Rather, the monodromy of the connection determined by A around the path φ can be thought of as an invertible linear transformation

$$\vec{v}_0 \rightarrow \mathbb{U}(\varphi)\vec{v}_0. \quad (13.57)$$

To be more explicit, let us take $n = 2$ and φ given by simple path with winding number 1. Say, for simplicity, it has lift $x(t) = t$. Suppose moreover that

$$A = \theta \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad (13.58)$$

is constant in x . Then the differential equation (13.41) is easily solved to give

$$\vec{v}(t) = \begin{pmatrix} \cos(\theta t) & \sin(\theta t) \\ -\sin(\theta t) & \cos(\theta t) \end{pmatrix} \vec{v}_0 \quad (13.59)$$

where $\vec{v}_0 \in \mathbb{R}^2$. Clearly, $\vec{v}(1)$ is not $\vec{v}_0$, in general.

Remark: Note that since $A(x)$ is invariant under $x \rightarrow x + 1$ it would make sense to define $\tilde{A}(z)$ as $A(x)$ for any x such that $z = \exp[2\pi i x]$. However, this is not the most convenient definition if we want the differential equation (13.41) to look the same in terms of z . Rather, if we define $\tilde{A}(z)$ so that $\tilde{A}(z)dz = A(x)dx$, that is, so that

$$\tilde{A}(z) := \frac{1}{2\pi i} e^{-2\pi i x} A(x) \quad (13.60)$$

then the equation (13.41) is equivalent to

$$\frac{d}{dt} \vec{v}(t) = \tilde{A}(z(t)) \frac{dz}{dt} \vec{v}(t) \quad \vec{v}(0) = \vec{v}_0. \quad (13.61)$$

That is, A should transform under change of coordinates as a 1-form.

Example 3: Let B be an open path-connected domain in $\mathbb{C}$. For example, B might be $\mathbb{C} - \{z_1, \dots, z_m\}$, i.e., $\mathbb{C}$ with some set of points deleted. We can also view it as the extended complex plane with the point at infinity also deleted: $B = \mathbb{CP}^1 - \{z_1, \dots, z_m, \infty\}$. Thus it is most definitely not simply connected for $m \geq 1$. Again let $E = B \times \mathbb{C}^n$ for some positive integer n . A path in B can be represented by $z(t)$. Our path lifting rule will be similar to Example 1: Choose a pair of $n \times n$ matrix-valued functions on B , call it $(A_z, A_{\bar{z}})$ where

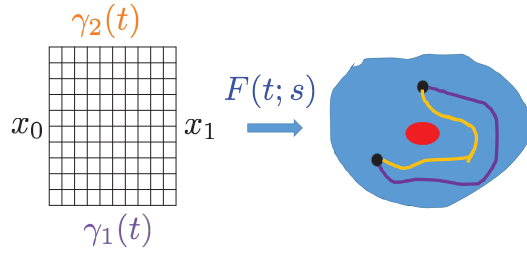


Figure 104: Illustrating the argument that for a flat connection on a domain in $\mathbb{C}$ the parallel transport only depends on the homotopy class of the curve with fixed endpoints. Using the homotopy, divide the region between the two curves into small regions by dividing the domain of the homotopy into sufficiently small squares. Then the monodromy around each small square is computed by the connection and its covariant derivatives: But these are all zero.

each matrix in the pair is a single-valued and nonsingular function of $(z, \bar{z})$.⁷⁷ Then the lifted path will be $(z(t), \vec{v}(t))$ where the differential equation is now:

$$\frac{d}{dt}\vec{v}(t) + \left(A_z(z(t), \bar{z}(t))\frac{dz}{dt} + A_{\bar{z}}(z(t), \bar{z}(t))\frac{d\bar{z}}{dt} \right) \vec{v}(t) = 0 \quad \vec{v}(0) = \vec{v}_0. \quad (13.62)$$

We can again “solve” the differential equation with the path-ordered exponential, and again there will be quite interesting monodromy. Note that we put both terms on the same side of the equation (thus A is related to the previous examples by a sign flip). This sign convention turns out to be more useful.

As an example of the monodromy let us consider a small loop based at z_0 , written as $z(t) = z_0 + \epsilon(t)$, where for fixed t , the complex number $\epsilon(t)$ will be taken to be small. In particular, the loop will be homotopically trivial, so there is a small disk D with basepoint z_0 such that ∂D is the image of φ . Then the leading nontrivial contribution to $\mathbb{U}(\varphi)$ will be appear at order $\mathcal{O}(\epsilon^2)$ and one can show that

$$\boxed{\mathbb{U}(\varphi) = 1 + \alpha F_{z\bar{z}}(z_0, \bar{z}_0) + \mathcal{O}(\epsilon^3)} \quad (13.63)$$

where

$$\alpha = \int_0^1 dt \dot{\epsilon}(t) \bar{\epsilon}(t) \quad (13.64)$$

is $((2i)$ times) the Euclidean area enclosed by the small loop at z_0 and

$$\boxed{F_{z\bar{z}} := \partial_z A_{\bar{z}} - \partial_{\bar{z}} A_z + [A_z, A_{\bar{z}}]} \quad (13.65)$$

One way to see that α is proportional to the area enclosed by the loop is to write

$$\alpha = \oint_{\varphi} d\epsilon \bar{\epsilon} = \int_D d\epsilon \wedge d\bar{\epsilon} \quad (13.66)$$

⁷⁷We are not assuming any relation between the complex conjugate $(A_z)^*$ and $A_{\bar{z}}$.

and the latter integral is $2i$ times the Euclidean area enclosed by the loop. One can show that in the full expansion of (13.63) in powers of $\epsilon(t)$ all the terms involve products of $\mathbb{F}_{z\bar{z}}$ and its (covariant) derivatives.

The expression $F_{z\bar{z}}$ is known as the *curvature of the connection*. It is more properly regarded as a locally matrix valued 2-form $F = F_{z\bar{z}}dz \wedge d\bar{z}$.

A particularly important class of connections are the *flat connections*, defined to be the connections with zero curvature. In this case, given a flat connection, *the monodromy matrix $\mathbb{U}(\wp)$ for a closed path only depends on the homotopy class of $\wp$ in B* . This is easy to show using (13.63) and the path composition property. From (13.63) the monodromy around a small loop must be trivial. But now if $F(t; s)$ is a homotopy from the closed loop $\wp_1(s)$ to $\wp_2(s)$ then we can divide up the square I^2 into many small squares, and the monodromy around each of these must be trivial, therefore, the monodromy around the full square must be trivial. Therefore $\mathbb{U}(\wp_1 \star \bar{\wp}_2) = 1$ and hence $\mathbb{U}(\wp_1) = \mathbb{U}(\wp_2)$.

In the special case that $A_{\bar{z}} = 0$ and A_z is a *holomorphic* function of z the property that $\mathbb{U}(\wp)$ only depends on the homotopy class can be seen more directly from the path-ordered exponential:

$$\mathbb{U}(\wp_t) = 1 + \int_{z_0}^{z(t)} A(z_1) dz_1 + \sum_{n=2}^{\infty} \int_{z_0}^{z(t)} dz_1 \int_{z_0}^{z_1} dz_2 \cdots \int_{z_0}^{z_{n-1}} dz_n A(z_1) \cdots A(z_n) \quad (13.67)$$

and now the assertion follows from Cauchy's theorem. Here dz_i is short for $\dot{z}(t_i)dt_i$ etc.

Remark: Many standard second order differential equations of mathematical physics can be understood as special cases of (13.62) in the holomorphic case. To make the connection note that the ODE

$$\frac{d^2}{dx^2}\psi + p(x)\frac{d}{dx}\psi + q(x)\psi = 0 \quad (13.68)$$

can be written as in equation (13.62) where

$$A_z = \begin{pmatrix} 0 & 1 \\ -q(z) & p(z) \end{pmatrix} \quad A_{\bar{z}} = 0 \quad (13.69)$$

where we assume that we can choose $t = x$ locally and $q(x)$ and $p(x)$ can be continued from some open domain in the real axis to analytic functions on some domain in the complex plane. Examples where this is the case include

1. *Schrödinger equation* :

$$p = 0 \quad q = \frac{2m}{\hbar^2}(V(z) - E) \quad (13.70)$$

(This assumes the potential energy $V(x)$ has an analytic extension to some domain in $\mathbb{C}$.)

2. *Bessel equation*

$$p(z) = \frac{1}{z} \quad q(z) = 1 - \frac{\nu^2}{z^2} \quad (13.71)$$

3. Gauss hypergeometric equation

$$p(z) = \frac{c - (1 + a + b)z}{z(1 - z)} \quad q(z) = \frac{-ab}{z(1 - z)} \quad (13.72)$$

4. confluent hypergeometric equation

$$p(z) = \frac{c - z}{z} \quad q(z) = \frac{-a}{z} \quad (13.73)$$

5. Legendre equation

$$p(z) = \frac{-2z}{1 - z^2} \quad q(z) = \frac{\nu(\nu + 1)}{1 - z^2} \quad (13.74)$$

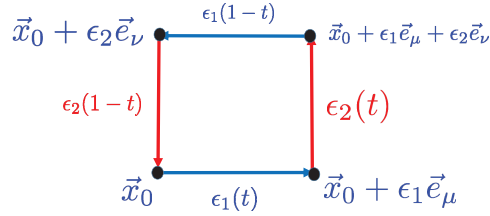


Figure 105: When ϵ_1 and ϵ_2 are small the entire contribution to the monodromy comes from the curvature of the connection, and the leading term is determined by the area of the loop times the curvature element in the plane spanned by the loop (in the tangent space) at $\vec{x}_0$.

Example 4: Now take B to be an open domain in $\mathbb{R}^m$ for any $m > 0$ and $E = B \times \mathbb{C}^N$, where m and N are in general completely unrelated. Choose coordinates x^μ , $\mu = 1, \dots, m$ on $\mathbb{C}^N$ and let $A_\mu(x)$ be a collection of m complex $N \times N$ matrix-valued functions on B . We assume they are single-valued and nonsingular. In close analogy to the above examples, this data suffices to define a connection on the fibration $p : E \rightarrow B$ given by projection on the first factor: Suppose $\varphi : [0, 1] \rightarrow B$ is a (piecewise differentiable) path in B from $\vec{x}_0$ to $\vec{x}_1$. Then we choose a lift $(\vec{x}_0, \vec{v}_0) \in E$ in the fiber above $\vec{x}_0$ and solve the differential equation

$$\frac{d}{dt} \vec{v}(t) + A_\mu(\vec{x}(t)) \frac{dx^\mu(t)}{dt} \vec{v}(t) = 0 \quad \vec{v}(0) = \vec{v}_0 \quad (13.75)$$

and $\mathbb{P}(\varphi)(t) = (\vec{x}(t), \vec{v}(t))$ is the lift. In general there will be interesting monodromy, already for small homotopically trivial paths near any point $\vec{x}_0$. The expressions are simple generalizations of (13.63) and (13.65). Indeed, any infinitesimal curve can be thought of as sitting in some plane passing through $\vec{x}_0$ and then it is simply a matter of changing back

from complex to real coordinates. Alternatively, we can consider a small path given by a composition of four open paths going around a square in the $x^\mu - x^\nu$ plane:

$$\begin{aligned}\wp_1(t) &= \vec{x}_0 + \epsilon_1(t)\vec{e}_\mu \\ \wp_2(t) &= (\vec{x}_0 + \epsilon_1\vec{e}_\mu) + \epsilon_2(t)\vec{e}_\nu \\ \wp_3(t) &= (\vec{x}_0 + \epsilon_2\vec{e}_\nu) + \epsilon_1(1-t)\vec{e}_\mu \\ \wp_4(t) &= \vec{x}_0 + \epsilon_2(1-t)\vec{e}_\nu\end{aligned}\tag{13.76}$$

and then $\wp = \wp_1 \star \wp_2 \star \wp_3 \star \wp_4$. Here $\vec{e}_\mu$ is a unit vector pointing in the x^μ direction and $\epsilon_i := \epsilon_i(t=1)$. See Figure 105. A simple computation shows that (13.63) and (13.65) are generalized to

$$\boxed{\mathbb{U}(\wp) = 1 - \epsilon_1\epsilon_2 F_{\mu\nu}(\vec{x}_0) + \cdots}\tag{13.77}$$

$$\boxed{F_{\mu\nu} := \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu]}\tag{13.78}$$

and the higher order terms in (13.77) are all of order $\epsilon_1^a \epsilon_2^b$ with $a > 0$ and $b > 0$ and $a + b > 2$.

Remark: It can be shown that the coefficients of the higher order terms in (13.77) are polynomials in $F_{\mu\nu}$ and its covariant derivatives in the μ and ν direction. In general, the covariant derivative of any matrix-valued function $\Phi(x)$ in the λ direction is

$$D_\lambda \Phi := \partial_\lambda \Phi + [A_\lambda, \Phi]\tag{13.79}$$

Exercise

Solve (13.62) for $B = \mathbb{C}^*$, $N = 1$, and $A_z = \frac{\mu}{z}$ and $A_{\bar{z}} = 0$, where μ is a complex number. Compute the monodromy of this connection around some simple closed curves in B .

Exercise

Let $x(t) = 3t$ and suppose that

$$A(x) = \begin{cases} -\alpha\sigma^3 & 0 \leq x \leq 1 \\ \beta\sigma^1 & 1 \leq x \leq 2 \\ \alpha\sigma^3 & 2 \leq x \leq 3 \end{cases}\tag{13.80}$$

where α and β are complex numbers.

Evaluate the path-ordered exponential and the exponential of the integral

$$\exp\left[\int_0^1 A(x(t))\dot{x}(t)dt\right]\tag{13.81}$$

and compare the answers.

Exercise

Give a careful derivation of equations (13.63), (13.65), (13.77), and (13.78).

Exercise

Show that if the matrices $A_\mu(x)$ are anti-hermitian, i.e. $(A_\mu(x))^\dagger = -A_\mu(x)$ then

$$\text{Pexp} \int_0^t A_\mu(x(t_1)) \frac{dx^\mu}{dt_1} dt_1 \quad (13.82)$$

is unitary.

Exercise *Gauge transformations*

Let $A(x)$ be a matrix-valued $n \times n$ complex matrix on $\mathbb{R}$ and $x \mapsto g(x)$ a differentiable map from $\mathbb{R}$ to $GL(n, \mathbb{C})$. Define a new matrix-valued function $\tilde{A}(x)$ by

$$A(x) = g(x)^{-1} \tilde{A}(x) g(x) + g(x)^{-1} \frac{d}{dx} g(x) \quad (13.83)$$

a.) Show that

$$d + \tilde{A} = g(x)(d + A)g(x)^{-1} \quad (13.84)$$

where $d = dx^\mu \frac{\partial}{\partial x^\mu} 1_{N \times N}$ is a first order differential operator and $A = dx^\mu A_\mu$.

b.) Show that, for any piecewise-differentiable path $x(t)$ from x_0 to x_1 we have

$$\text{Pexp} \left[- \int_0^1 \tilde{A}(x(t)) \dot{x}(t) dt \right] = g(x_1) \text{Pexp} \left[- \int_0^1 A(x(t)) \dot{x}(t) dt \right] g(x_0)^{-1} \quad (13.85)$$

c.) Show by direct computation that, if $\tilde{F}_{\mu\nu}$ is computed from $\tilde{A}_\mu$ then

$$F_{\mu\nu}(x) = g(x)^{-1} \tilde{F}_{\mu\nu}(x) g(x) \quad (13.86)$$

d.) Show that the commutator of matrix-valued first order differential operators gives the curvature:

$$[D_\mu, D_\nu] = F_{\mu\nu} \quad (13.87)$$

Use this to give another proof of the gauge transformation rule of part (c).

e.) Suppose $\Phi(x)$ and $\tilde{\Phi}(x)$ are matrix valued functions of x^μ related by $\tilde{\Phi}(x) = g(x)\Phi(x)g(x)^{-1}$. Show by direct computation that

$$\tilde{D}_\lambda \tilde{\Phi}(x) = g(x) D_\lambda \Phi(x) g(x)^{-1} \quad (13.88)$$

where $\tilde{D}_\lambda$ is the covariant derivative computed with $\tilde{A}_\lambda$.

f.) Show that $[D_\mu, \Phi] = D_\mu \Phi$. Use this to give another proof of the gauge transformation rule in (e).

13.6 Solution of the lifting problem for covering spaces

If $p : \widehat{X} \rightarrow X$ is a covering then we can give a very satisfying and complete solution to the lifting problem:

Theorem: Let $\widehat{p} : \widehat{X} \rightarrow X$ be a covering map. If Y is a path connected space then a map $\bar{f} : Y \rightarrow X$ has a lifting $f : Y \rightarrow \widehat{X}$ if and only if

$$\bar{f}_*(\pi_1(Y, y_0)) \subset \widehat{p}_*(\pi_1(\widehat{X}, \hat{x}_0)) \quad (13.89)$$

Idea of proof: One direction is trivial: If $\bar{f}$ has a lifting f then $\widehat{p} \circ f = \bar{f}$ and hence the inclusion (13.89) holds.

The other direction is harder: Suppose we have (13.89). Then we attempt to construct the lift f by lifting paths. Choose a basepoint $y_0 \in Y$ and suppose we want to define the lift at some other point $y \in Y$. Choose a path γ from y_0 to y . Then $\bar{f} \circ \gamma$ is a path in X from $x_0 = \bar{f}(y_0)$ to $x = \bar{f}(y)$. Now, choose some $\hat{x}_0$ in the fiber over x_0 . We will define our lift so that $f(y_0) = \hat{x}_0$. Next, using the property that $\widehat{p}$ is a covering map we lift the path $\bar{f} \circ \gamma$ to a path $\hat{\gamma}$ in $\widehat{X}$ with initial point $\hat{x}_0$. We want to define $f(y) = \hat{\gamma}(1)$. The obvious potential problem with this definition is that $\hat{\gamma}(1)$ might very well depend on the choice of γ . If this were the case then $f : Y \rightarrow \widehat{X}$ would not be well-defined.

What could go wrong in our definition? Suppose we choose two paths γ_1 and γ_2 connecting y_0 to y . (In general they will not be homotopic.) Then we lift $\bar{f} \circ \gamma_1$ to $\hat{\gamma}_1$ and $\bar{f} \circ \gamma_2$ to $\hat{\gamma}_2$, both beginning at $\hat{x}_0$. It is not completely obvious that they have the same endpoints. That is, we want to show that $\hat{\gamma}_1(1) = \hat{\gamma}_2(1)$. However, the closed path $\gamma_1 \star \gamma_2$ based at y_0 in Y maps to the closed path $\bar{f} \circ (\gamma_1 \star \gamma_2)$ based at x_0 in X and, by (13.89) we know there is a closed path $\rho \in \Omega_{\hat{x}_0}(\widehat{X})$ such that $\widehat{p} \circ \rho$ is homotopic to $\bar{f} \circ (\gamma_1 \star \gamma_2)$. That homotopy has a lift, since the lift of $\widehat{p} \circ \rho$ is obviously ρ . Therefore, the closed loop $\bar{f} \circ (\gamma_1 \star \gamma_2)$ has a lift with *no nontrivial monodromy*. Put differently, the unique lift of $\bar{f} \circ (\gamma_1 \star \gamma_2)$ which is $\hat{\gamma}_1 \star \hat{\gamma}_2$ must be a closed curve. In particular $\hat{\gamma}_1$ and $\hat{\gamma}_2$ must have the same endpoints in $\widehat{X}$, and hence $f : Y \rightarrow \widehat{X}$ is well-defined.

The map constructed this way is a lift of $\bar{f}$, by construction. With some straightforward further arguments one can show that f constructed in this way is indeed continuous. ♠

Exercise

Take $X = Y = \widehat{X} = S^1$. Let $p : \widehat{X} \rightarrow X$ be the n -fold covering $p(z) = z^n$. Which maps $\bar{f} : S^1 \rightarrow S^1$ have a lift to $\widehat{X}$?

13.7 The universal cover

If X is path connected and $p : \widehat{X} \rightarrow X$ is a covering map then from (13.39) we know that $p_* : \pi_1(\widehat{X}, \hat{x}_0)$ is an injection into $\pi_1(X, x_0)$. In general, the image will be a proper subgroup. In other words, taking a covering of a space “dilutes” the fundamental group.

One can in fact dilute the fundamental group to the trivial group. Up to a suitable notion of equivalence (discussed below) there is a unique such covering space called the *universal covering space* of X . We will denote it by $\tilde{X}$.

Theorem Every nice ⁷⁸ path-connected topological space X has a connected and simply connected covering space $\tilde{X}$.

Proof: Choose a basepoint x_0 . We will construct the universal cover out of the path fibration $p : (\mathcal{P}(X, x_0), \alpha_0) \rightarrow (X, x_0)$, where α_0 is the constant path at x_0 . Recall that $p(\alpha) = \alpha(1)$, is the endpoint of the path.

We define $\tilde{X}$ by imposing an equivalence relation, denoted $\sim_U$ on $\mathcal{P}(X, x_0)$. We say that two paths α, β based at x_0 are equivalent, $\alpha \sim_U \beta$ if they have the same endpoints and are homotopy equivalent with fixed endpoints. That is $\alpha \sim_U \beta$ if $\alpha(1) = \beta(1)$ and $\alpha \cong \beta$ relative to $\{0, 1\}$. The universal cover of X is defined to be the set of equivalence classes under this relation:

$$\tilde{X} := \mathcal{P}(X, x_0) / \sim_U \quad (13.90)$$

We take the compact-open topology on $\mathcal{P}(X, x_0)$ and the quotient topology on $\tilde{X}$. If α is a based path in X then we let $U(\alpha) := [\alpha]_{\sim_U}$ denote the equivalence class under $\sim_U$.

Note there is an obvious projection inherited from that of the path fibration:

$$\tilde{p} : \tilde{X} \rightarrow X \quad \tilde{p}(U(\alpha)) := \alpha(1) \quad (13.91)$$

Now, one can show that (13.91) is indeed a covering map. This is intuitively clear: If $U(\alpha)$ ends at a point $x = \alpha(1)$ then a small neighborhood of $U(\alpha)$ in $\tilde{X}$ is made by concatenating small curves based at x . That is, choose an open neighborhood V of x , then

$$\{U(\alpha \star \eta) : \eta : [0, 1] \rightarrow V\} \quad (13.92)$$

forms an open set around $U(\alpha)$. Since we quotient by homotopy equivalence, for small neighborhoods V this set will be homeomorphic to V . Therefore $\tilde{p}$ is a covering map.

What is the fiber above a point x_0 ? The fiber of the path fibration was the based loop space $\Omega_{x_0}(X)$. Since we mod out by homotopy equivalence to get to $\tilde{X}$, the fiber $\tilde{X}_{x_0}$ of $\tilde{X} \rightarrow X$ above x_0 is in 1-1 correspondence with the elements of $\pi_1(X, x_0)$. Note that there is a distinguished point $\tilde{x}_0 = U(\alpha_0) \in \tilde{X}$ over x_0 , corresponding to the homotopy class of the constant loop. That is, $\tilde{x}_0$ is the point in $\tilde{X}$ representing the homotopically trivial based loops at x_0 .

What about the fiber over some other point $x_1 \neq x_0$? Choosing a homotopy class of paths from $x_0 \rightarrow x_1$ and applying equation (12.5) in Section §12.2 above we see that the fiber $\pi_1(X, x_1)$ above x_1 can be put in one-one correspondence with the fiber $\pi_0(X, x_0)$, albeit not in a canonical way.

⁷⁸See below for a discussion of “nice.”

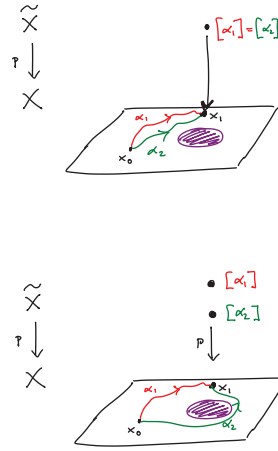


Figure 106: In (a) the paths are identified in $\tilde{X}$: $[\alpha_1] = [\alpha_2]$. In (b) $[\alpha_1]$ and $[\alpha_2]$ define two distinct points on the cover, even though the endpoints are the same, and hence $p([\alpha_1]) = p([\alpha_2]) = x_1$.

Now we would like to show that $\tilde{X}$ is connected and simply connected. To show that it is connected consider an arbitrary point $\tilde{x} \in \tilde{X}$. Then $\tilde{x} = U(\alpha)$ can be represented by some path α based at x_0 . Now we define a path $\varphi : [0, 1] \rightarrow \tilde{X}$ by $\varphi(t) = U(\alpha_t)$ where α_t is the based path at x_0 defined by $s \mapsto \alpha(st)$, $0 \leq s \leq 1$. The path φ connects $\tilde{x}_0$ to $\tilde{x}$.

Now we use the LES of a fibration:

$$0 \rightarrow \pi_1(\tilde{X}, \tilde{x}_0) \rightarrow \pi_1(X, x_0) \rightarrow \pi_0(\mathcal{F}) \rightarrow \pi_0(\tilde{X}) \rightarrow \pi_0(X) \rightarrow 0 \quad (13.93)$$

We have just shown that $\pi_0(\tilde{X})$ consists of a single point. Moreover, we also showed that $\pi_0(\mathcal{F}) \cong \pi_1(X, x_0)$. Since the kernel of $\pi_0(\tilde{X}) \rightarrow \pi_0(X)$ is zero it follows that the homomorphism $\pi_1(X, x_0) \rightarrow \pi_0(\mathcal{F})$ is an isomorphism and hence $\pi_1(\tilde{X}, \tilde{x}_0) = \{1\}$ is the trivial group. ♠

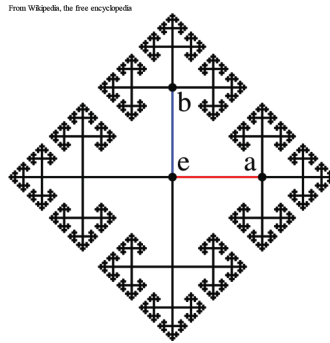


Figure 107: The universal cover of the the figure eight is the Cayley graph for the free group on two generators.

Examples

1. The universal cover of S^1 is $\mathbb{R}$.
2. The universal cover of S^n is S^n , for $n > 1$. (Note that although the path space $\mathcal{P}(X, x_0)$ is contractible, the universal cover $\tilde{X}$ is in general not contractible, as this example shows.)
3. The universal cover of $\mathbb{R}P^n$ is $\mathbb{R}$, for $n = 1$ and is S^n for $n > 1$.
4. The universal cover of an n -dimensional torus is $\mathbb{R}^n$.
5. The universal cover of a figure eight (i.e. the wedge of two circles) is shown in Figure 107. The path from e to a is the lift of a path going around one (say, the first) circle with a certain orientation and the horizontal path in the other direction is the lift of the path going around the first circle in the other direction. Similarly, the two vertical segments from e are the lifts of paths around the second circle of winding number one with opposite orientations.
6. Let us consider the universal covers of surfaces with $\chi = 0$. The universal cover of a sphere with two boundaries is (i.e. a cylinder) can be taken to be $\mathbb{R} \times [0, 1]$. The projection map can be thought of as the projection for the equivalence relation $(x, y) \sim (x + 1, y)$. The universal cover of a Mobius strip is also $\mathbb{R} \times [0, 1]$: Now the equivalence relation is $(x, y) \sim (x + 1, 1 - y)$. The universal cover of T^2 is $\mathbb{R}^2$, as noted above. The universal cover of the Klein bottle is again $\mathbb{R}^2$, but now with the equivalence relation generated by identifying $(x, y) \sim (L - x, y + \beta)$ and $(x, y) \sim (x + L, y)$.⁷⁹
7. The universal cover of a topological surface with $\chi < 0$ is the upper half-plane. (This is a nontrivial theorem.)

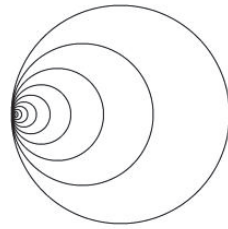


Figure 108: The Hawaiian Earring does not have a simply connected universal cover.

⁷⁹Here L and β are positive real numbers. For purposes of topology they can be set to 1 (or any other positive real number). However, they have significance when we include metrics, and indeed have physical significance in applications to string theory and orientifolds.

Remarks

1. We should say more about what we mean by a “nice” path-connected space X . Technically, it is a “semilocally simply connected space.” Any space you are likely to meet in a physics problem will satisfy this condition. The technical condition says that around every point $x \in X$ there is a neighborhood U of x so that the inclusion map $\iota_* : \pi_1(U, x) \rightarrow \pi_1(X, x)$ is trivial. That is, every based loop at x inside U is null homotopic. CW complexes and topological manifolds are semi-locally simply connected. An example of a topological space that does not satisfy this criterion is the so-called “Hawaiian earring” of Figure 108.
2. Why is it called a universal cover? We claim that if $\hat{p} : \hat{X} \rightarrow X$ is *any* cover, then there is also a covering $p' : \tilde{X} \rightarrow \hat{X}$ so that

$$\begin{array}{c} \tilde{X} \\ \downarrow p' \\ \hat{X} \\ \downarrow \hat{p} \\ X \end{array} \quad (13.94)$$

gives the universal covering with $\tilde{p} = \hat{p} \circ p'$. To prove this we simply note that since $\pi_1(\tilde{X}) = 0$ we have a unique solution to the lifting problem:

$$\begin{array}{ccc} & \hat{X} & \\ & \downarrow \hat{p} & \\ \tilde{X} & \xrightarrow{\tilde{p}} & X \end{array} \quad (13.95)$$

Because $\tilde{p}$ and $\hat{p}$ are covering maps and $\tilde{p} = \hat{p} \circ p'$, by restricting to suitable neighborhoods and using the local inverse to $\hat{p}$ we see that p' is a local homeomorphism, so it is also a covering map.

There is a sense in which the universal cover is “unique.” But to make that precise we need a suitable notion of equivalence of covers. This is provided by

Definition A *morphism* of two covering spaces is a map φ such that we have a commutative diagram.

$$\begin{array}{ccc} \hat{X}_1 & \xrightarrow{\varphi} & \hat{X}_2 \\ & \searrow p_1 & \swarrow p_2 \\ & X & \end{array} \quad (13.96)$$

An *isomorphism* φ is a morphism such that there exists a morphism $\psi : \hat{X}_2 \rightarrow \hat{X}_1$ so that $\varphi \circ \psi$ and $\psi \circ \varphi$ are the identity maps.

Now we can prove uniqueness of the universal cover: If $(\tilde{X}_1, \tilde{p}_1)$ and $(\tilde{X}_2, \tilde{p}_2)$ are two candidate universal covers then we have

$$\begin{array}{ccccc}
 & & & \tilde{X}_1 & \\
 & & \nearrow \psi & \downarrow \tilde{p}_1 & \\
 & \tilde{X}_2 & \xrightarrow{\tilde{p}_2} & X & \\
 \nearrow \varphi & \downarrow \tilde{p}_2 & & \downarrow = & \\
 \tilde{X}_1 & \xrightarrow{\tilde{p}_1} & X & \xrightarrow{=} & X
 \end{array} \tag{13.97}$$

Look at the outer triangle. By uniqueness of the lifting for covers we see that $\psi \circ \varphi = \text{Id}_{\tilde{X}_1}$. The other direction is similar.

Remark: When we constructed the universal cover we made a *choice* of basepoint $x_0 \in X$. It is interesting to see how the universal cover depends on this choice, and how that is compatible with the uniqueness we have just explained. Suppose we made another choice of basepoint $x'_0 \in X$. Choose a path γ from x_0 to x'_0 . Then there is a map

$$\Phi_\gamma : \mathcal{P}(X, x'_0) \rightarrow \mathcal{P}(X, x_0) \tag{13.98}$$

given by

$$\Phi_\gamma : \alpha \mapsto \gamma \star \alpha \tag{13.99}$$

Note that $p \circ \Phi_\gamma = p'$ since α and $\gamma \star \alpha$ have the same endpoint. Moreover, this map descends to a continuous map

$$\varphi_\gamma : \mathcal{P}(X, x'_0) / \sim_U \rightarrow \mathcal{P}(X, x_0) / \sim_U \tag{13.100}$$

defined by

$$\varphi_\gamma(U(\alpha)) = U(\gamma \star \alpha). \tag{13.101}$$

Note, in particular, that this map only depends on the homotopy class of γ (with fixed endpoints), so we can write $\varphi_{[\gamma]}$. Moreover, it is clear that $\varphi_{[\bar{\gamma}]}$ is a left and right inverse to $\varphi_{[\gamma]}$. Therefore, $\varphi_{[\gamma]}$ is an isomorphism of covers.

More generally, if γ_1 is a path from x_0 to x'_0 and γ_2 is a path from x'_0 to x''_0 then

$$\varphi_{[\gamma_1 \star \gamma_2]} = \varphi_{[\gamma_1]} \circ \varphi_{[\gamma_2]} \tag{13.102}$$

Therefore, if γ_1, γ_2 are two paths from x_0 to x'_0 then in general $\varphi_{[\gamma_1]}$ will be different from $\varphi_{[\gamma_2]}$ if the homotopy classes $[\gamma_1]$ and $[\gamma_2]$ are different. Indeed

$$\varphi_{[\gamma_1]} \circ (\varphi_{[\gamma_2]})^{-1} = \varphi_{[\gamma_1 \star \bar{\gamma}_2]} = \varphi_{[\rho]} \tag{13.103}$$

where ρ is the based loop $\gamma_1 \star \bar{\gamma}_2$ at x_0 . This is the automorphism of $\tilde{X}$ defined by

$$U(\alpha) \rightarrow U(\rho \star \alpha). \tag{13.104}$$

We will return to this formula. ⁸⁰

Some sources:

Armstrong,

Massey, *Algebraic Topology: An Introduction*

Munkres,

A. Hatcher, *Algebraic Topology*

Exercise

Consider the covering $\hat{p} : S^1 \rightarrow S^1$ given by $\hat{p}(z) = z^n$ for a nonzero integer n . Show that the map p' in equation (13.94) is given by

$$p'(x) = e^{2\pi i x/n} \quad (13.105)$$

Exercise

We noted above that the universal cover of the Klein bottle is $\mathbb{R}^2$. Show that there is an intermediate cover (13.94) where $\hat{X}$ is the torus, and $\hat{p}$ is a double cover.

13.8 The Galois correspondence between covers of X and subgroups of $\pi_1(X)$

There is a very beautiful geometric analog of the Galois correspondence in the subject of field theory (“field theory” in the sense of abstract algebra, not physics).

Theorem Let X be a path connected space with a universal cover. There is a 1-1 correspondence between (isomorphism classes of) path-connected covers of X and (isomorphism classes of) subgroups of $\pi_1(X)$.

Proof: Choose a basepoint $x_0 \in X$. This pins down a definite group, namely, $\pi_1(X, x_0)$ in the isomorphism class $\pi_1(X)$. Then the Galois correspondence is a 1-1 correspondence between isomorphism classes of covers and conjugacy classes of subgroups of $\pi_1(X, x_0)$.

One direction is fairly easy: Given a cover $\hat{p} : \hat{X} \rightarrow X$ choose a basepoint $\hat{x}_0$ in the fiber over x_0 and define

$$H := \hat{p}_* \left(\pi_1(\hat{X}, \hat{x}_0) \right) \subset \pi_1(X, x_0) \quad (13.106)$$

This gives a definite subgroup H , but it depended on a choice of $\hat{x}_0$ in the covering space. What happens if we make another choice $\hat{x}'_0$? Since, by assumption, $\hat{X}$ is path connected

⁸⁰A sophisticated way to say all this is the following: The universal covers of a connected space form a groupoid, and there is an equivalence of categories between the fundamental groupoid of X and the groupoid of universal covers.

there is a path ζ in $\widehat{X}$ from $\hat{x}_0$ to $\hat{x}'_0$. Now recall the discussion of equation (12.5). We have an isomorphism

$$\psi_\zeta : \pi_1(\widehat{X}, \hat{x}_0) \rightarrow \pi_1(\widehat{X}, \hat{x}'_0) \quad (13.107)$$

given by

$$\psi_\zeta : [\hat{\alpha}] \mapsto [\bar{\zeta} \star \hat{\alpha} \star \zeta] \quad (13.108)$$

where $\hat{\alpha}$ is a loop in $\widehat{X}$ based at $\hat{x}_0$. Now note that $\hat{p}(\zeta)$ is a closed path in X based at x_0 . It follows that

$$H' := \hat{p}_* \left(\pi_1(\widehat{X}, \hat{x}'_0) \right) \subset \pi_1(X, x_0) \quad (13.109)$$

is related to H by conjugation:

$$H' = g_0^{-1} H g_0 \quad (13.110)$$

where $g_0 = [\hat{p}(\zeta)]$ is a homotopy class in X based at x_0 .

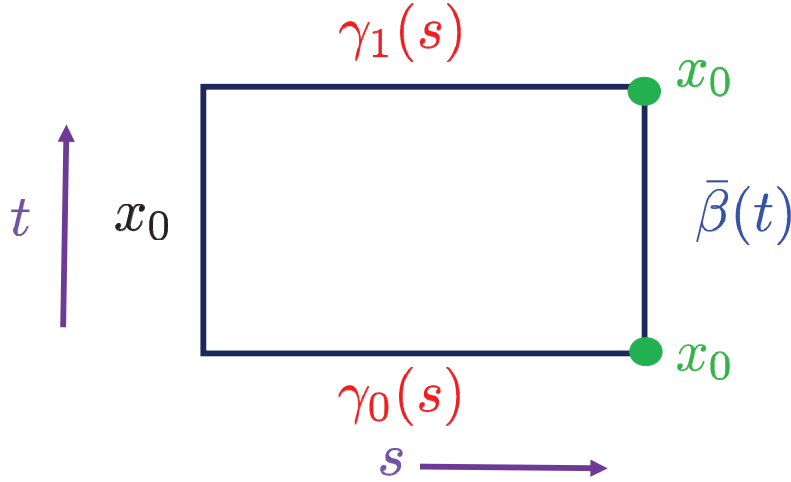


Figure 109: A loop in X_H based at $[\alpha_0]_H$ defines a map $\beta(t; s)$ from the square into X .

The other direction takes more work. Suppose now we are given a specific subgroup $H \subset \pi_1(X, x_0)$. We need to produce a corresponding cover. We will do this by taking a quotient of the universal cover by an equivalence relation $\sim_H$. We say that $U(\alpha_1) \sim_H U(\alpha_2)$ if they have the same endpoints, and if the homotopy class of the closed path $\alpha_1 \star \bar{\alpha}_2$ in X based at x_0 is in H .

Notice this condition makes sense as an equivalence relation on the path space, and we could have simply defined $\sim_H$ that way. From that point of view, the universal cover is just given by the equivalence relation $\sim_H$ where H is the trivial subgroup. Now for $\sim_H$

we are considering some paths in Figure 106(b) to be equivalent points, even though they were inequivalent in $\tilde{X}$. Let us denote the equivalence class $[U(\alpha)]_{\sim_H}$ simply $[\alpha]_H$.

Now we define

$$X_H := \tilde{X} / \sim_H \quad (13.111)$$

and $p_H([\alpha]_H) = \alpha(1)$ is again the endpoint of any representative path α in X . It is not difficult (but tedious) to show that p_H is a covering map.

We claim that

$$(p_H)_*(\pi_1(X_H, [\alpha_0]_H)) = H. \quad (13.112)$$

Let us first show that the LHS of this equation is even a subgroup of H . Consider a loop in X_H based at $[\alpha_0]_H$. We can represent it by a loop $t \mapsto [\beta(t)]_H$, $0 \leq t \leq 1$. This is a loop of paths, so there is a single function $\beta(t; s)$ such that the path at fixed t is $s \mapsto \beta(t; s)$, $0 \leq s \leq 1$. The path at fixed t is not necessarily a closed path in X . However, $\beta(t; 0) = x_0$ since we constructed X_H using paths beginning at x_0 . Moreover, $t \mapsto p_H([\beta(t)]_H) = \beta(t; 1) := \tilde{\beta}(t)$ must be a closed path in X based at x_0 , and hence $\beta(1; 1) = \beta(0; 1) = x_0$. See Figure 109. In this figure $s \mapsto \gamma_0(s)$ and $s \mapsto \gamma_1(s)$ are based loops in X at x_0 and must represent the basedpoint $[\alpha_0]_H$. That is, they are based loops at X whose homotopy class is in H . It then follows from Figure 109 that the closed loop $\tilde{\beta}$ in X has homotopy class

$$[\tilde{\beta}] = [\tilde{\gamma}_0 \star \gamma_1] = [\gamma_0]^{-1} \cdot [\gamma_1] \in H \quad (13.113)$$

Thus $(p_H)_*$ indeed maps $\pi_1(X_H, [\alpha_0]_H)$ into H . It is of course a homomorphism. Since p_H is a covering map we know that $(p_H)_*$ is injective. Here is an explicit argument that it is injective: It follows from (13.113) that if $[\tilde{\beta}] = 1$ then it can be homotoped to the constant map at x_0 . When this is done the paths $s \mapsto \beta_t(s)$ are closed loops in X based at x_0 and Figure 109 is just a homotopy between γ_0 and γ_1 and in particular $[\gamma_0] = [\beta_t] = [\gamma_1]$. Therefore the loop $t \mapsto [\beta_t]_H$ is just the constant map $[\alpha_0]_H$.

Finally, we need to show that $(p_H)_*$ is surjective. Suppose that $h \in H$ is represented by a closed path α based at x_0 . Then we let $\wp_\alpha$ be the *closed* path in X_H given by $\wp_\alpha(t) = [\alpha_t]_H$ where $\alpha_t : [0, 1] \rightarrow X$ is the *open path* defined by $\alpha_t(s) := \alpha(ts)$, $0 \leq s \leq 1$. By construction $p_H([\alpha_t]_H) = \alpha(t)$. Therefore

$$(p_H)_*([\wp_\alpha]) = [\alpha] = h. \quad (13.114)$$



Remark: Note that in this classification of coverings it is crucial that we assume $\hat{X}$ is path connected. If D is any set with discrete topology then, rather trivially, $\hat{X} = X \times D$, with p being the projection on the first factor, is a covering. In general D has nothing to do with $\pi_1(X)$ or any of its subgroups. But such coverings are disconnected.

Example Let us construct the connected covers of $X = S^1$. We know that $\pi_1(S^1, 1) \cong \mathbb{Z}$. The subgroups H of $\mathbb{Z}$ are $N\mathbb{Z}$ where we can take N to be a positive integer. Then the

corresponding cover $p_H : X_H \rightarrow X$ is $X_H = S^1$ with $p_H(z) = z^N$. We have already encountered these covers. Now we know that this exhausts all the connected covers.

Exercise

Consider $p : S^1 \rightarrow S^1$ defined by $p(z) = z^{-N}$ where N is a positive integer. Show that this is one of the covers described above.

13.8.1 Galois correspondence and normal subgroups

Now let us return to the morphisms between covers and consider the automorphisms or self-isomorphisms of a cover $\hat{p} : \hat{X} \rightarrow X$. These are homeomorphisms $\varphi : \hat{X} \rightarrow \hat{X}$ such that the diagram

$$\begin{array}{ccc} \hat{X} & \xrightarrow{\varphi} & \hat{X} \\ & \searrow \hat{p} & \swarrow \hat{p} \\ & X & \end{array} \quad (13.115)$$

commutes. The automorphisms of a cover form a group, known as the *covering group* or the *group of Deck transformations*. We denote it by $\mathcal{D}(\hat{X}, \hat{p})$. Often the projection is understood and we just write $\mathcal{D}(\hat{X})$.

Theorem Suppose $H \subset \pi_1(X, x_0)$ is a subgroup and $p_H : X_H \rightarrow X$ is the corresponding cover constructed in (13.111). Then

$$\mathcal{D}(X_H, p_H) \cong N(H)/H \quad (13.116)$$

where $N(H)$ is the normalizer of H in $\pi_1(X, x_0)$.

Proof: Suppose $g_0 \in N(H)$. Choose a representative loop γ based at x_0 so $g_0 = [\gamma]$. Then there is a continuous map

$$\Phi_\gamma : \mathcal{P}(X, x_0) \rightarrow \mathcal{P}(X, x_0) \quad (13.117)$$

given by

$$\Phi_\gamma : \alpha \mapsto \gamma \star \alpha \quad (13.118)$$

If $g_0 \in N(H)$ this map descends to

$$\varphi_{[\gamma]} : [\alpha]_H \mapsto [\gamma \star \alpha]_H \quad (13.119)$$

If $g_0 \notin N(H)$ then $\varphi_{[\gamma]}$ might well not be well-defined.

Another way to see that g_0 in fact must be in the normalizer of H is the following: Note that $\varphi_{[\gamma]}$ maps the basepoint $[\alpha_0]_H$ of X_H to some other point $[\gamma]_H$ in the fiber of p_H over x_0 . Now recall the discussion of how a change of basepoint $\hat{x}_0$ conjugates the image group under $\hat{p}_*$. In particular, $[\gamma_t]_H$ is a path between $[\alpha_0]_H$ and $[\gamma]_H$ and $p \circ \varphi = p$ implies that

$$H = g_0^{-1} H g_0 \quad (13.120)$$

Thus we have a map, in fact a homomorphism, $\Psi : N(H) \rightarrow \mathcal{D}(X_H)$,

$$\Psi : [\gamma] \mapsto \varphi_{[\gamma]}. \quad (13.121)$$

It should be clear that the kernel of Ψ is exactly H .

Now we need to see that Ψ is in fact surjective. Suppose $\varphi : X_H \rightarrow X_H$ is a Deck transformation. Then it maps the basepoint $[\alpha_0]_H$ of X_H to some other point in the fiber above x_0 , that is:

$$\varphi([\alpha_0]_H) = [\gamma]_H \quad (13.122)$$

where γ is some closed loop in X based at x_0 . In general the element $[\gamma] \in \pi_1(X, x_0)$ need not be in H .

Now, for any $\alpha \in \mathcal{P}(X, x_0)$ we have

$$p_H(\varphi([\alpha_t]_H)) = p_H([\alpha_t]_H) \quad (13.123)$$

since $p_H \circ \varphi = p_H$. As usual $\alpha_t : s \mapsto \alpha(ts)$. But this means that $t \mapsto \varphi([\alpha_t]_H)$ is a lift of $\alpha(t)$. But the basepoint of the lift is $\varphi([\alpha_0]_H) = [\gamma]_H$. Therefore, by uniqueness of the path lifting:

$$\varphi([\alpha_t]_H) = [\gamma \star \alpha]_H \quad (13.124)$$

Therefore $\varphi = \varphi_{[\gamma]}$ and hence Ψ is surjective. Thus we have

$$1 \rightarrow H \rightarrow N(H) \xrightarrow{\Psi} \mathcal{D}(X_H) \rightarrow 1 \quad (13.125)$$

thus completing the proof. ♠

In particular, if H is a *normal* subgroup then $N(H) = \pi_1(X, x_0)$ and hence the covering group for X_H is the quotient group $\pi_1(X, x_0)/H$. For the case of the universal cover $H = \{1\}$ is the trivial subgroup and therefore $\mathcal{D}(\tilde{X}) = \pi_1(X, x_0)$. The fact that the fiber $\mathcal{F}$ is itself a group leads to an important new idea, described in the next section.

Example Let us return to the example of $X = S^1$ with the inclusion $H \subset \pi_1(X)$ isomorphic to $N\mathbb{Z} \subset \mathbb{Z}$, where we can take N to be a positive integer. Then the corresponding cover $p_H : X_H \rightarrow X$ is $X_H = S^1$ with $p_H(z) = z^N$. Now, π_1 is abelian so $N(H) = \pi_1$, so the group of Deck transformations of X_H is isomorphic to $\mathbb{Z}/N\mathbb{Z}$. Indeed, if ω is an N^{th} root of unity then $\phi_\omega : X_H \rightarrow X_H$ given by $\phi_\omega(z) = \omega z$ is a Deck transformation.

Remarks:

1. There is a deep analogy between the theory of covering spaces and Galois theory of field extensions in number theory. If we have a polynomial equation with integer coefficients

$$a_n x^n + \cdots + a_1 x + a_0 = 0 \quad (13.126)$$

then we can define a field extension F of $\mathbb{Q}$ by adjoining the complex roots. The group of Deck transformations, which permutes the sheets, is analogous to the Galois

group, which permutes the roots. We should think of $\mathbb{Q}$ as analogous to the field of functions on X and F as analogous to the field of functions on $\widehat{X}$. Note that there are certainly “more” functions on $\widehat{X}$ since any function on X pulls back to a function on $\widehat{X}$, but there might be some functions on $\widehat{X}$ not of this form. The analog of the universal cover is the field $\overline{\mathbb{Q}}$ of all algebraic numbers. The different subfields of F covering $\mathbb{Q}$ are analogous to the other covering spaces intermediate between the universal cover and $\mathbb{Q}$.

Exercise

Show that if $g_0 \in N(H)$ then equation (13.119) is indeed well-defined. ⁸¹

13.9 Coverings and principal bundles with discrete structure group

We now present a different viewpoint about the universal cover and the associated coverings X_H . This viewpoint makes use of the ideas of group actions and principal bundles.

Let X be a path-connected topological space with a universal cover. Choose a basepoint $x_0 \in X$ and a corresponding universal cover as constructed above. We have seen that the group of Deck transformation $\mathcal{D}(\widetilde{X})$ is isomorphic to $\pi_1(X, x_0)$. Moreover, if $g \in \pi_1(X, x_0)$ and g is represented by $g = [\gamma]$ then the corresponding Deck transformation is

$$\varphi_{[\gamma]} : U(\alpha) \mapsto U(\gamma \star \alpha) \quad (13.127)$$

Now, an easy computation shows that

$$\varphi_{[\gamma_1]} \circ \varphi_{[\gamma_2]} = \varphi_{[\gamma_1] \cdot [\gamma_2]} \quad (13.128)$$

Thus, the action of the group of Deck transformations on $\widetilde{X}$ is a (left) group action of $\pi_1(X, x_0)$ on $\widetilde{X}$.

We can also form a right group action of $\pi_1(X, x_0)$ on $\widetilde{X}$ by using $\varphi_{[\bar{\gamma}]}$, and this will be a little more convenient so, for $g = [\gamma] \in \pi_1(X, x_0)$ we define:

$$\tilde{x} \cdot g := \varphi_{[\bar{\gamma}]}(\tilde{x}) \quad (13.129)$$

Thanks to the definition of a covering space this action is properly discontinuous and hence *the universal cover of X is a principal Γ bundle over X with $\Gamma = \pi_1(X, x_0)$.*

Examples Let us go through our previous list of examples of universal covers from this viewpoint:

⁸¹ *Answer:* The map is potentially ill-defined because $[\rho \star \alpha]_H = [\alpha]_H$ where ρ is a closed loop based at x_0 with $[\rho] \in H$. So we must show that $[\gamma \star \rho \star \alpha]_H = [\gamma \star \alpha]_H$. But this is true iff the homotopy class in X of the based loop $\gamma \star \rho \star \alpha \star \bar{\alpha} \star \bar{\gamma}$ is in H . But this is homotopic to $\gamma \star \rho \star \bar{\gamma}$. Since this must hold for every ρ with $[\rho] \in H$ it follows that $[\gamma]$ must be in the normalizer of H in $\pi_1(X, x_0)$.

1. The universal cover of $B = S^1$ is $P = \mathbb{R}$ with structure group $\mathbb{Z}$ and action $x \cdot n := x + n$.
2. For $n > 1$ the universal cover of $B = \mathbb{RP}^n$ is $P = S^n$ with structure group $\mathbb{Z}_2$ and action $\vec{x} \cdot \sigma = -\vec{x}$.
3. The universal cover of an n -dimensional torus $B = T^n$ is $\mathbb{R}^n$ with structure group $\mathbb{Z}^n$ and action $\vec{x} \cdot \vec{n} := \vec{x} + \vec{n}$.
4. The universal cover of a figure eight (i.e. the wedge of two circles) is shown in Figure 107. The structure group is the free group on two generators g_1, g_2 . The free right action is defined as follows: g_1 moves a vertex to the next vertex one step to the right, while g_1^{-1} moves it one step to the left. Similarly g_2 moves a vertex to the next vertex one step up and g_2 moves it one step down. The action on the points in the intervals joining the vertices is similar.
5. The description of the equivalence relation on the universal covers of the surfaces with $\chi = 0$ in fact generated group actions. The universal cover of a sphere with two boundaries is (i.e. a cylinder) can be taken to be $\mathbb{R} \times [0, 1]$. The projection map can be thought of as the projection for the equivalence relation $(x, y) \sim (x + 1, y)$. The universal cover of a Mobius strip is also $\mathbb{R} \times [0, 1]$: Now the equivalence relation is $(x, y) \sim (x + 1, 1 - y)$. The universal cover of T^2 is $\mathbb{R}^2$, as noted above. The universal cover of the Klein bottle is again $\mathbb{R}^2$, but now with the equivalence relation generated by identifying $(x, y) \sim (L - x, y + \beta)$ and $(x, y) \sim (x + L, y)$.
6. The universal cover of a topological surface with $\chi < 0$ is the upper half-plane. (This is a nontrivial theorem.)

DISCUSS FUNDAMENTAL GROUP OF THE KLEIN BOTTLE IN DETAIL AS PRINCIPAL BUNDLE STRUCTURE GROUP.

Now suppose that $H \subset \Gamma$ is a subgroup. Then we can consider the quotient space $\tilde{X}/H$. This still has a projection to X . It will be a fiber bundle with fiber space Γ/H . Since Γ and H are discrete groups Γ/H has discrete topology and hence $\pi_j(\Gamma/H) = 0$ for $j > 0$. Applying the exact homotopy sequence we have

$$1 \rightarrow \pi_1(\tilde{X}/H, [\alpha_0]) \xrightarrow{p_*} \Gamma \xrightarrow{\partial_1} \Gamma/H \rightarrow 1 \quad (13.130)$$

where, for general H the last map is just a map of pointed sets. Thus we conclude that $\pi_1(\tilde{X}/H, [\alpha_0]) \cong H$. In particular we can identify

$$X_H \cong \tilde{X}/H \quad (13.131)$$

Now, in general X_H is *not* a principal bundle, because the fibers Γ/H are homogeneous spaces, but are not groups.

Now note that there is a well-defined *right* action of the group $N(H)$ on the set of *left*-cosets Γ/H : For $n \in N(H)$ and any $g \in \Gamma$ we can define

$$\varphi(n, gH) := (gn)H \quad (13.132)$$

(Check that this is a right action, and that it is only well-defined for $n \in N(H)$!!) Moreover, if $n \in H$ the action is trivial. This gives another way of understanding why the group of Deck transformations of X_H is $N(H)/H$.

DISCUSS MONODROMY FOR UNIQUE CONNECTION ON PRINCIPAL BUNDLES WITH DISCRETE STRUCTURE GROUP. The point is now we get a homomorphism $\pi_1(X) \rightarrow \Gamma$. Now write classification of principal Γ bundles. So compare classification of coverings and principal bundles with discrete group. They are different things.

13.10 Branched covers and multi-valued functions

Let e be a positive integer. The map $p : \mathbb{C}^* \rightarrow \mathbb{C}^*$ defined by $p(z) = z^e$ is a covering map. If we write w for the coordinate on the target $\mathbb{C}^*$ and z for the coordinate on the domain $\mathbb{C}^*$ then we can express the map as

$$w = z^e \quad (13.133)$$

It is tempting to rewrite this as

$$z = w^{1/e} \quad (13.134)$$

but of course the function $w^{1/e}$ is not single-valued on the complex w -plane, although it can be *locally* defined as an analytic function. The correct point of view is that $w^{1/e}$, while not locally defined, is in fact a perfectly well-defined function, but on a *different space* related to the original w -space as a covering space.

This viewpoint leads to a very fruitful way of looking at functions defined implicitly by algebraic equations. They are often not single-valued, but become so when pulled back to an appropriate covering space.

Before explaining further let us note that the map p can be extended to a nice holomorphic map $p : \mathbb{C} \rightarrow \mathbb{C}$. However, it is now no longer a covering map: There is no neighborhood U of $w = 0$ so that $p^{-1}(U)$ is a disjoint union of neighborhoods homeomorphic to U . Indeed, for a covering map the fiber above different points always has the same homotopy type. But for p the fiber above $w = 0$ consists of the single point $z = 0$, while the fiber above any $w \neq 0$ consists of e distinct points. This map is an example of a *branched covering*. The point $w = 0$ on the base space is called a *branch point*. The point $z = 0$ on the covering space is called a *ramification point*. The positive integer e is called the *ramification index*. When working with branched covers it is common to use the term *unramified covering* to mean what we have simply been calling a “covering.”

In general, a *branched covering* is a map of pairs $\pi : (Y, R) \rightarrow (X, B)$ where $R \subset Y$ and $B \subset X$ are of real codimension two. R is called the *ramification locus* and B is called the *branch locus*. The map $\pi : Y - R \rightarrow X - B$ is an unramified covering and if this unramified covering is an n -fold covering we say the branched covering is an n -fold branched cover. On the other hand, near any point $b \in B$ there is a neighborhood U of b and local coordinates

$$(x_1, \dots, x_{d-2}; w) \in \mathbb{R}^{d-2} \times \mathbb{C}, \quad (13.135)$$

where $\dim_{\mathbb{R}} X = d$ and $w = 0$ describes the branch locus $B \cap U$. More importantly, $\pi^{-1}(U) = \coprod_{\alpha} \tilde{U}_{\alpha}$ is a disjoint union of neighborhoods in Y of points $r_{\alpha} \in R$ with local

coordinates

$$(x_1, \dots, x_{d-2}; \xi_\alpha) \in \mathbb{R}^{d-2} \times \mathbb{C} \quad (13.136)$$

so that the map $\pi_\alpha : \tilde{U}_\alpha \rightarrow U$ (where π_α is just the restriction of π) is just given by

$$\pi_\alpha : (x_1, \dots, x_{d-2}; \xi_\alpha) \rightarrow (x_1, \dots, x_{d-2}; \xi_\alpha^{e_\alpha}) \quad (13.137)$$

where e_α are positive integers called *ramification indices*.

In plain English: For any $b \in B$ there are several points r_α in the preimage of π above b and near any r_α the map π looks like a mapping of unit disks in the complex plane $\xi \rightarrow w = \xi^e$. Note that for an n -fold covering

$$\sum_\alpha e_\alpha = n \quad (13.138)$$

The case where exactly one ramification index is $e = 2$ and all the others are equal to one is called a *simple branch point*.

Remarks

1. Branched covers often arise in algebraic geometry. Here we have a holomorphic map of algebraic varieties $\pi : Y \rightarrow X$ which is generically a covering. However, there is a complex codimension one subvariety $R \subset Y$ and $B \subset X$ where the map is not quite a covering. At generic smooth points we can find local holomorphic coordinates $(\xi, z_2, z_3, \dots)$ near a point on R so that R is given locally by the equation $\xi = 0$ and local holomorphic coordinates $(\eta, z_2, z_3, \dots)$ near B so that B is given locally by $\eta = 0$, and the map π is just $\pi : \xi \rightarrow \eta = \xi^e$. The number e is known as the *ramification index*.
2. A useful invariant of a branched cover is expressed in terms of π_1 . Note that $X - B$ is in general not simply connected because B has real codimension two. Choose a basepoint $x_0 \in X - B$, and label the preimages in $p^{-1}(x_0)$ by $x_0^{(1)}, \dots, x_0^{(n)}$ in some definite order. Then the lift of closed curves based at x_0 defines a homomorphism $\pi_1(X - B, x_0) \rightarrow S_n$ called the monodromy representation of the cover.

13.10.1 Example: Hyperelliptic curves

A very rich example, which we will return to later is given by the set of points:

$$\Sigma = \{(x, y) | y^2 = P(x)\} \subset \mathbb{C}^2 \quad (13.139)$$

where $P(x)$ is a polynomial. We will make the very important assumption that all of its roots are *simple*, i.e. $P'(\rho) \neq 0$ for every root ρ of P .

We claim that $\pi : \Sigma \rightarrow \mathbb{C}$ given by $\pi(x, y) = x$ is a branched cover of degree 2, with branch locus B equal to the set of roots of P and $R = \{(\rho_i, 0) | \rho_i \in B\} \subset \Sigma$.

If x is not a root of the polynomial then there is a $2 : 1$ covering space around some open neighborhood of x . The covers are distinguished by the two values of y . Now suppose

ρ_i is a root. There is a neighborhood $\mathcal{U}_i$ around ρ_i and a 1-1 analytic redefinition of functions x, y on $\pi^{-1}(\mathcal{U}_i)$ so that the map π becomes the map $\xi \rightarrow \eta = \xi^2$.

To see this first let $\eta = P'(\rho_i)(x - \rho_i)$. This is plainly a 1-1 invertible analytic map! (We should write η_i , but we do not do so to avoid cluttering the formulae.) Now we have

$$y^2 = \eta + b_2\eta^2 + \cdots + b_n\eta^n \quad (13.140)$$

so $\eta \sim y^2$ when η is small. Therefore, for small η we can invert equation (13.140) to produce a *convergent* power series

$$\eta = y^2 + c_4y^4 + c_6y^6 + \cdots \quad (13.141)$$

Now define $\xi(y)$ as a power series (again convergent):

$$\xi = y (1 + c_4y^2 + c_6y^4 + \cdots)^{1/2} = y + d_3y^3 + d_5y^5 + \cdots \quad (13.142)$$

The analytic redefinition $y \rightarrow \xi(y)$ given by (13.142) is the one we seek. Note that indeed on Σ we have $\eta = \xi^2$.

It is easy to see how to invert the power series (13.142). Suppose we start again with (13.140) and now let ξ be one of the two roots of $\xi^2 = \eta$. Observe that we can solve (13.140) if

$$y = \xi (1 + b_2\xi^2 + b_3\xi^4 + \cdots + b_n\xi^{2n-2})^{1/2} \quad (13.143)$$

Of course, the two roots of (13.140) for a fixed η are given by using (13.143) with the two roots $\pm\xi$ of $\eta = \xi^2$. Now we can expand the squareroot in (13.143) as a *convergent* power series. This defines the analytic coordinate redefinition $\xi \rightarrow y$ so that the projection map becomes $\pi : \xi \rightarrow \eta = \xi^2$. Of course (13.143) and (13.142) are inverse power series.

Figure 110: (a.) The monodromy representation describing the covering of $\Sigma - R \rightarrow \mathbb{C} - B$. Around each of the curves γ_i the monodromy is the nontrivial element of $\mathbb{Z}_2$. (b.) Therefore, the monodromy around a curve surrounding two branch points ρ_i, ρ_j is trivial.

If we remove the branch points then we have an unramified covering $\pi : \Sigma - R \rightarrow \mathbb{C} - B$. If we choose a basepoint $x_0 \in \mathbb{C} - B$ then we can compute the monodromy representation. We choose a system of simple closed curves surrounding each of the roots ρ_i . The monodromy transformation around each one is the permutation of the two sheets.

As opposed to unramified covers, which dilute $\pi_1(X, x_0)$, branched covers can create new elements of $\pi_1(E, e_0)$.

NEED TO REALIZE THAT WE HAVE A TORUS.

13.10.2 Riemann-Hurwitz formula

Riemann-Hurwitz formula

Classify branched covers with specified branch locus.

14. CW Complexes

Most “reasonable” spaces admit triangulations, simplicial decompositions, and cell-decompositions. Cell decompositions are a little more flexible. Such decompositions can be useful in computing topological invariants involving homotopy, homology, and cohomology.

The basic idea is to build up spaces from elementary building blocks by gluing. One useful implementation of this idea is the use of *CW complexes*.⁸²

Recall that the closed n -disk is defined to be:

$$D^n := \{\vec{x} \in \mathbb{R}^n \mid \vec{x} \cdot \vec{x} \leq 1\} \quad (14.1)$$

Its interior is

$$\text{Int}(D^n) := \{\vec{x} \in \mathbb{R}^n \mid \vec{x} \cdot \vec{x} < 1\} \quad (14.2)$$

For $n = 0$, we take $\text{Int}(D^0) = D^0 = pt$. We now define a space homeomorphic to D^n to be a *closed n -cell* and a space homeomorphic to $\text{Int}(D^n)$ to be an *open n -cell*.

Definition: Given a Hausdorff topological space X a *CW decomposition* is a disjoint decomposition

$$X = \coprod_{\alpha} e_{\alpha} \quad (14.3)$$

such that

1. Each e_{α} is an open cell of dimension $n(\alpha) \geq 0$.
2. To each α is a continuous map

$$f_{\alpha} : D^{n(\alpha)} \rightarrow X \quad (14.4)$$

such that f maps $\text{Int}(D^n)$ homeomorphically to e_{α} .

3. Each $x \in \overline{e_{\alpha}} - e_{\alpha}$ belongs to a cell e_{β} of strictly lower dimension $n(\beta) < n(\alpha)$.

Remarks:

1. There are actually cases in which it is important to allow an infinite number of cells, and the dimension of the cells can be unbounded. In this case there are further technical conditions for defining a CW complex.
2. Note that $\overline{e_{\alpha}}$ need not be a closed n -cell because there are identifications on the boundary of $D^{n(\alpha)}$.

⁸²We are following Milnor and Stasheff, ch. 6 and Bredon, *Topology and Geometry*, Ch. IV.8.

3. The name “CW” is unhelpful to any but the most specialized. It is short for “closure-finite weak,” and was invented by J.H.C. Whitehead in 1949.

Example 1: *A cell decomposition of S^n :* We think of S^n as the unit sphere in $\mathbb{R}^{n+1}$. Then we consider two cells:

$$\begin{aligned} e_0 &= \{(0, \dots, 0, -1) \in S^n\} \\ e_n &= \{(x_1, \dots, x_{n+1}) \in S^n \mid x_{n+1} > -1\} \end{aligned} \quad (14.5)$$

For e_0 we take the trivial map f_0 . For e_n we take the map

$$f_n : D^n \rightarrow S^n \quad (14.6)$$

which maps the interior of D^n to e_n and maps ∂D^n to the point e_0 . To be precise

$$f_n : \vec{x} \mapsto (\hat{x} \sin(\pi|\vec{x}|), \cos(\pi|\vec{x}|)) \quad (14.7)$$

Example 2: *Another cell decomposition of S^n :* Now we have two cells in each dimension up to n . We think of them as subspaces of S^n defined by

$$\begin{aligned} e_{i,+} &= \{(x_1, \dots, x_{i+1}, 0, \dots, 0) \in S^n \mid x_{i+1} > 0\} \\ e_{i,-} &= \{(x_1, \dots, x_{i+1}, 0, \dots, 0) \in S^n \mid x_{i+1} < 0\} \end{aligned} \quad (14.8)$$

Now we define

$$f_{i,\pm} : D^i \rightarrow S^n \quad (14.9)$$

by

$$\begin{aligned} f_{i,+}(\vec{x}) &\mapsto (\hat{x} \sin(\frac{\pi}{2}|\vec{x}|), \cos(\frac{\pi}{2}|\vec{x}|)) \\ f_{i,-}(\vec{x}) &\mapsto (\hat{x} \sin(\frac{\pi}{2}|\vec{x}|), -\cos(\frac{\pi}{2}|\vec{x}|)) \end{aligned} \quad (14.10)$$

The advantage of this decomposition is that it is “equivariant” that is, it is clearly compatible with the transformation $x \rightarrow -x$ on S^n taking a point to its antipodal. It is compatible in the sense that the cells are mapped into each other.

Example 3: *A cell decomposition for $\mathbb{R}P^n$:*

A CW decomposition leads to the important notion of an *n-skeleton*. We can build up the space in successive dimensions by attaching cells.

We begin with the “0-skeleton,” X_0 this is simply the disjoint union of 0-cells. Then to get X_1 we consider the closure of the union of all the 1-cells and 0-cells. To get X_j we take the closure of the union of all j -cells with X_{j-1} . Note that

$$X_j/X_{j-1} = \bigvee_{\alpha: n(\alpha)=j} S^j \quad (14.11)$$

is a one-point union of spheres of dimension j , one for each j -dimensional cell.

Now suppose $n(\alpha) = j + 1$. Then

$$f_\alpha : D_\alpha^{j+1} \rightarrow X_{j+1} \quad (14.12)$$

In particular we can restrict f_α to the boundary $\partial D_\alpha^{j+1} \cong S^j$ to produce a map:

$$f_\alpha : S^j \rightarrow X_{j+1} \quad (14.13)$$

Now choose any cell e_β with $n(\beta) = j$. Then we can consider the composite:

$$S^j \xrightarrow{f_\alpha} X_{j+1} \xrightarrow{\pi} X_j / X_{j-1} \cong \bigvee_{\gamma: n(\gamma)=j} S_\gamma^j \xrightarrow{\pi_\beta} S_\beta^j \quad (14.14)$$

This composite map is denoted by $f_{\alpha,\beta}$. Note that it is just a map of a j -sphere to a j -sphere. The topology of the space X is large encoded in these maps. It is one of the reasons why understanding the homotopy of maps of spheres to spheres is important in topology.

For example, $n(\alpha, \beta) := \deg(f_{\alpha,\beta})$ can be used to define a chain complex from which one can compute the homology and cohomology of the space X .

Exercise

a.) Show that in the cell-decomposition (14.8) et. seq. we have

$$\partial e_j^+ = e_{j-1}^+ + e_{j-1}^- \quad (14.15)$$

$$\partial e_j^- = -e_{j-1}^+ - e_{j-1}^- \quad 1 \leq j \leq n \quad (14.16)$$

$$\partial e_0^\pm = 0 \quad (14.17)$$

where the $+$ means union and the $-$ means union with the opposite orientation.

b.) Check that $\partial^2 = 0$.

Exercise

Let $f : S^1 \rightarrow S^1$ be the n -fold covering defined by $f(z) = z^n$ where z is a complex number of modulus one. Consider the space $M_n = D^2 \cup_f S^1$. Show that

$$\pi_1(M_n) = \mathbb{Z}/n\mathbb{Z}$$

Exercise

Show that *every* map $f : X \rightarrow Y$ is homotopy equivalent to an inclusion! That is, there always exists a space $\tilde{Y}$ homotopy equivalent to Y so that the composition $f : X \rightarrow \tilde{Y}$ is an inclusion map.

14.1 The Euler character

Let N_k be the number of k -cells of a CW decomposition of X . Clearly this is not a topological invariant: subdivision yields homeomorphic cell complexes.

But – instead we can define:

$$\chi(X) := \sum_{k \geq 0} (-1)^k N_k$$

Examples

1. From our first cell-decomposition of spheres we see that there is one cell in dimension n and one in dimension 0, and hence

$$\chi = 1 + (-1)^n = \begin{cases} 0 & n \text{ odd} \\ 2 & n \text{ even} \end{cases} \quad (14.18)$$

2. From our second cell decomposition of S^n we get

$$\chi = 2 - 2 + 2 - 2 \pm \cdots + (-1)^n 2 = \begin{cases} 0 & n \text{ odd} \\ 2 & n \text{ even} \end{cases} \quad (14.19)$$

3. From our cell decomposition of $\mathbb{R}P^n$ we similarly get

$$\chi = 1 - 1 + 1 - 1 \pm \cdots + (-1)^n 1 = \begin{cases} 0 & n \text{ odd} \\ 1 & n \text{ even} \end{cases} \quad (14.20)$$

4. From our cell-decomposition of surfaces with boundary we get:

$$\chi = 2 - 2g - b - c \quad (14.21)$$

One can show:

1. The Euler character does not depend on the CW decomposition of X .
2. The Euler characteristic is a homotopy invariant (hence, a homeomorphism invariant), but is not a cobordism invariant: Note that any orientable Riemann surface can be filled in and is hence cobordant to zero.

15. Bordism and Cobordism

16. Counting solutions of an equation: The degree of a map

Many of the spectacular applications of quantum field theory to topology and geometry eventually come down to “counting” solutions to nonlinear differential equations defining holomorphic curves, instantons, monopoles, etc. in field theory. In this section we describe some very elementary examples of this, as background to the general idea.

Figure 111: Graph of $x^2 - t$. Counting the number of solutions depends on t .

Suppose we have a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ that depends on control parameters t , call it $f(x; t)$. For example we could take $n = 1$, $t \in \mathbb{R}$ and

$$f(x; t) = x^2 - t \quad (16.1)$$

In general we might want to count the number of solutions to the equation

$$f(x; t) = 0 \quad (16.2)$$

Of course, the number of solutions is locally constant in t . Nevertheless, it is not a topological invariant, and depends on the sign of t . This is illustrated in our simple example in 111.

Nevertheless, we *can* get a topological invariant by computing a signed sum.

We will restrict attention to *proper* maps, so that $f^{-1}(K)$ is compact if K is compact. Thus the inverse image of a point $q \in \mathbb{R}^n$ is compact. Moreover we should restrict to counting solutions

$$f(p; t) = q \quad (16.3)$$

to q 's which are *regular values* of f .

Definition:

a.) A *critical point* of a map $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$ is a point p such that $df : T_p \mathbb{R}^m \rightarrow T_q \mathbb{R}^n$ has rank smaller than n , that is, it is not surjective.

b.) A *critical value* of f is the image of a critical point.

c.) A point $q \in \mathbb{R}^n$ which is not a critical point is a *regular value*. Put differently, a regular value of a map $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$ is a point q such that $df : T_p \mathbb{R}^m \rightarrow T_q \mathbb{R}^n$ is onto for *each* of the preimages. We say f is transverse at q . Warning: A point q for which the set of inverse images is empty is considered to be a regular value.

An important theorem about critical and regular values is

Sard's Theorem: If $f : M^m \rightarrow \mathbb{R}^n$ is a smooth map then the set of critical values has measure zero.

We will not prove this. But we note an immediate and useful corollary:

Theorem: If $f : M^m \rightarrow S^n$ is smooth and $m < n$ then it is homotopic to the constant map.

Proof: By Sard's theorem there must be a point $q \in S^n$ which is not in the image of f (since any point in the image of f is a critical point). But $S^n - \{q\}$ can be contracted to a point. Call this contraction c_t . Then $c_t \circ f$ is a homotopy to the constant map. ♠

Next note that the preimage of a regular value of a proper map $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a finite set of points. In this case one good answer to how to count solutions to $f(p) = q$ is given by the following theorem:

Theorem. Let q be a regular value for f . Then the signed sum

$$\sum_{p \in f^{-1}(q)} \text{signdet}\left(\frac{\partial f^i}{\partial x^j}\right)|_p \quad (16.4)$$

is independent of the choice of q and is a topological invariant under smooth deformations of the function f . It is called the *degree of f* .

Proof: Using the fact that q is a regular value of f there is a neighborhood $\mathcal{U}$ around q so that each $p \in f^{-1}(q)$ has a neighborhood where f is a diffeomorphism. Now choose an n -form α with support in $\mathcal{U}$ which integrates to $+1$. (Such a form is called a “bump form.”) Now the following integral is then easily evaluated:

$$\deg(f) := \int_{\mathbb{R}^n} f^*(\alpha) = \sum_{p \in f^{-1}(q)} \text{signdet}\left(\frac{\partial f^i}{\partial x^j}\right)|_p \quad (16.5)$$

Now, the bump form α represents a generator of $H_{\text{cpt}}^n(\mathbb{R}^n) \cong \mathbb{Z}$. Therefore, any two choices of α differ by

$$\alpha_1 - \alpha_2 = d\eta \quad (16.6)$$

where η is compactly supported. Moreover, any two choices of q also differ by d of a compactly supported form η . Therefore, since f is proper, $f^*(\eta)$ is compactly supported and $f^*(\alpha_1) - f^*(\alpha_2) = df^*(\eta)$, we can use Stokes theorem and the value of the integral is unchanged. ♠

Remarks:

- Note that this theorem is a very special case of Thom's ideas on bordism: The oriented bordism class of the inverse image is invariant.

-

We can generalize this to the case where f is transverse to a subspace of $\mathbb{R}^n$.

Figure 112: Graph of a function f . Counting solutions with signs is invariant.

Figure 113: The graph of the function f is not transverse to the zero section

Figure 114: The relative homology class in the unit interval bundle over S^1 does not have well-defined self-intersection with itself, because the intersection point can move off to the boundary. It does have well-defined intersection with the zero-section.

16.1 Intersection interpretation

Now look at the graph of f in 111:

$$\Gamma(f) = \{(x, f(x))\} \tag{16.7}$$

We can regard this as a section of an oriented vector bundle $\mathbb{R} \rightarrow \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ and ask about the intersection number with the zero section

$$Z = \{(x, 0)\} \tag{16.8}$$

Note that

$$\deg(f) = \langle \Gamma(f), Z \rangle \tag{16.9}$$

In topological field theory one generalizes this idea to arbitrary (oriented) vector bundles over oriented manifolds. To do so we require the theory of the Thom class.

There are two pitfalls to keep in mind:

- It can happen that the zeroes of f are not isolated as in 113. This typically happens in topological field theory.

- One must be careful defining intersection numbers in noncompact spaces, or even in spaces with boundary. A simple example of one of the subtleties is illustrated in 114.

16.2 The degree for proper maps between manifolds

The above discussion generalizes to arbitrary compact oriented manifolds.

Suppose $f : M_1 \rightarrow M_2$ is a map of compact oriented n -manifolds. Then $H^n(M_1; \mathbb{Z}) \cong \mathbb{Z}$ and $H^n(M_2; \mathbb{Z}) \cong \mathbb{Z}$. Let ω_1, ω_2 be integral generators. Then the degree of f is defined by:

$$f^*(\omega_2) = (\deg f)\omega_1 \quad (16.10)$$

We claim this is the same as:

$$f_*([M_1]) = (\deg f)[M_2] \quad (16.11)$$

This just follows from Poincaré duality. To tie this to the previous discussion choose a regular value $q \in M_2$ of f and a ball $B_q \subset M_2$ such that

$$\overline{f^{-1}(B_q)} \subset \cup_{p \in f^{-1}(q)} D_p \quad (16.12)$$

where p are regular points and f is a diffeomorphism from a ball B_p around each p to B_q . The picture is:

Figure 115: f is a diffeomorphism to B_q on each of the small balls, although it might be orientation preserving or reversing on each ball.

Now, choose a Poincaré dual α to the point $q \in M_2$, $\eta(q \rightarrow M_2)$. Once again, we can take it to be a bump form supported in a neighborhood of p and once again its cohomology class is the class of a generator of $H^n(M_2; \mathbb{Z}) \cong \mathbb{Z}$.

Therefore, we have

$$\int_{[M_1]} f^*(\omega_2) = \sum_{p \in f^{-1}(q)} \int_{[M_1]} f^*\alpha = \sum_{p \in f^{-1}(q)} \text{signdet}(df)|_p \quad (16.13)$$

Note in the last equation that $df_p : T_p M_1 \rightarrow T_p M_2$ and to define the determinant we must choose bases. Since M_1 and M_2 are oriented the sign is in fact unambiguous and doesn't depend on which oriented basis we choose.

16.3 Examples

1. Of course, the winding number $f : S^1 \rightarrow S^1$ is an example. $f(\theta) = e^{i\psi(\theta)}$, then

$$\begin{aligned} \deg(f) &= \int_{S^1} f^* \left(\frac{d\psi}{2\pi} \right) \\ &= \int \frac{d(\log f)}{2\pi i} \\ &= \oint \frac{f'(z)}{f(z)} \frac{dz}{2\pi i} \end{aligned} \tag{16.14}$$

In the last line we have interpreted $S^1 \subset \mathbb{C}$ as the unit circle and, by Fourier analysis have decomposed $f(z) = \sum a_n z^n$ as an analytic function in a neighborhood of S^1 .

2. Suppose we stereographically project $S^2 \rightarrow \mathbb{C}$ and consider $f : S^2 \rightarrow S^2$ as a map $\mathbb{C} \rightarrow \mathbb{C}$. Then it is easy to show that

$$\deg(f) = \frac{i}{2\pi} \int_{\mathbb{C}} \frac{df \wedge d\bar{f}}{(1 + |f|^2)^2} \tag{16.15}$$

There is another useful way to say this. Consider a map

$$\vec{n} : \mathbb{R}^2 \rightarrow S^2 \tag{16.16}$$

subject to a boundary condition

$$\lim_{x \rightarrow \infty} \vec{n}(x) = \vec{n}_0 \tag{16.17}$$

so that we can regard $\vec{n} : S^2 \rightarrow S^2$.

Now

$$\omega = \frac{1}{8\pi} \epsilon^{ijk} x^i dx^j \wedge dx^k|_{S^2} \tag{16.18}$$

and

$$\begin{aligned} \deg(\vec{n}) &= \int_{S^2} \vec{n}^*(\omega) \\ &= \int_{\mathbb{R}^2} \vec{n}^*(\omega) \\ &= \frac{1}{8\pi} \int_{\mathbb{R}^2} \epsilon^{ijk} \epsilon^{\alpha\beta} n^i \partial_\alpha n^j \partial_\beta n^k d\xi^1 d\xi^2 \end{aligned} \tag{16.19}$$

is an integer.

Exercise

Write an explicit unit volume form for S^n and therefore write an integral formula for $\deg(f)$. Also write this using stereographic projection $S^n \rightarrow \mathbb{R}^n$.

Exercise

Compute the degree of $f : x \rightarrow -x$ on S^n

Exercise

Show that every map $f : S^n \rightarrow T^n$ has degree zero.

Exercise

- a.) Consider the map $f : SU(2) \rightarrow SU(2)$ given by $f(U) = U^k$. Compute the degree of f .
- b.) What happens for $f : SU(n) \rightarrow SU(n)$ given by $f(U) = U^k$?
-

16.4 Computing $\pi_n(S^n)$

Figure 116: A map $f : S^n \rightarrow S^n$ can be homotoped to this form: It is a sum of maps $S^n \rightarrow S^n$ which represent the identity map with winding number ± 1 . The signed sum of these windings is the degree of the map.

First consider the 0-dimensional sphere. It has two points so $\pi_0(S^0) = \mathbb{Z}_2$. Choosing one of the points as a basepoint we can define $\pi_k(S^0)$ for $k > 0$. Clearly, $\pi_k(S^0) = 0$ for $k > 0$.

Next:

Theorem For $n > 0$

$$\begin{aligned}\pi_k(S^n, p_0) &= \{0\} & 0 \leq k < n \\ \pi_n(S^n, p_0) &= \mathbb{Z}\end{aligned}\tag{16.20}$$

Sketch of Proof: For a proof see Bott and Tu, pp. 214-215:

1. First one shows that we can take f to be differentiable.
2. For $k < n$ we use the Corollary of Sard's theorem, explained above, that any smooth map $f : S^k \rightarrow S^n$ is homotopic to the constant map.
2. For $k = n$ some nontrivial arguments are needed. We have already shown that we map associate to a smooth map $f : S^n \rightarrow S^n$ its *degree*.

$$\deg(f) := \int_{S^n} f^*(\omega) \quad (16.21)$$

where $\int_{S^n} \omega = 1$ is the unit volume form. The identity map has winding number 1. We claim that this gives the required isomorphism. The nontrivial part consists in showing that, up to homotopy the map is equivalent to a sum (in the sense of homotopy groups) of maps of winding number ± 1 as in 116. This is done in Bott and Tu. ♠

16.5 The degree as a “topological field theory integral”

At this point we can introduce a basic characature of how some topological field theories lead to interesting topological invariants.

Suppose $s^a(\phi^1, \dots, \phi^m)$ with $a = 1, \dots, n$ is a map of real variables $(\phi^1, \dots, \phi^m) \in \mathbb{R}^m$, which are to be thought of as “zero dimensional fields.” The $s^a(\phi)$ are to be thought of as “equations whose solutions we want to count.” For example, in real applications they might be the equations for self-dual instantons, or magnetic monopoles, or holomorphic maps.

For each of the “fields” ϕ^i we introduce an anticommuting partner ψ^i and for each of the equations introduce anticommuting variables χ_a , $a = 1, \dots, n$.

Then by doing the fermionic integral and then the bosonic integral one can show that

$$\deg(s) = -\left(\frac{1}{\sqrt{2\pi\hbar}}\right)^n \int_{\widehat{\mathbb{R}^n}} \prod_i d\phi^i d\psi^i \int_{\Pi(\mathbb{R}^n)^*} \prod_a d\chi_a e^{-\frac{1}{2\hbar} s^a(\phi) s^a(\phi) + i\chi_a \frac{ds^a}{d\phi^j} \psi^j} \quad (16.22)$$

The integral should be regarded as a “zero-dimensional path integral” and the topological invariance is nicely explained by a supersymmetry of the action of the form:

$$\begin{aligned} Q\phi^i &= \psi^i \\ Q\psi^i &= 0 \\ Q\chi_a &= -\frac{i}{\hbar} s^a \end{aligned} \quad (16.23)$$

Q is also defined to be odd and it satisfies the Leibniz rule. This defines the action of Q on any analytic superfield.

17. Overview of the uses of topology in field theory

1. soliton sectors
2. classifying phases of matter by order parameters

3. instanton sectors
4. topological terms in actions
5. framework for a formulation of field theory
6. computation of topological invariants using field theory

17.1 Digression: Physics and the classification of manifolds

18. Solitons and soliton sectors

Sometimes simple energetic considerations impose nontrivial topologies on spaces of fields. There can then be disconnected components of fieldspace where field configurations in different components are separated by an “infinite energy barrier.” Here we give a brief overview of some of the general ideas.

18.1 Soliton sectors

Let us assume spacetime is foliated by equal time slices, so

$$\mathcal{S}_d = \mathbb{R} \times \mathcal{S}_{d-1} \quad (18.1)$$

We aim to classify the connected components of configuration space. In general we have some collection of maps to a target manifold M_n .

Note that time-evolution is a *continuous* process, so that, among other things, time development always defines a homotopy of the field configurations for fixed t . If space is compact then we can simply use the compact-open topology on the space of fields and the field configurations falls into disconnected components labelled by

$$\pi_0(\text{Map}(\mathcal{S}_{d-1}, M_n)) = [\mathcal{S}_{d-1}, M_n] \quad (18.2)$$

Example: One very significant example is closed string theory which is based on a two-dimensional theory and the target M_n is the spacetime in which the particles of string theory move and interact. Here $\mathcal{S}_{d-1} = S^1$. Then, if $\pi_1(M_n) \neq 0$ there are *winding sectors* of the Hilbert space of the first-quantized string theory. This is an important feature not present in point particle theories, where $\mathcal{S}_{d-1} = pt..$

When we quantize the theory, these different sectors of the configuration space give different sectors of the Hilbert space of states: By the correspondence principle, the field theoretic *soliton configurations* should correspond to *quantum states* in the theory. They are often different superselection sectors distinguished by a topological conservation law.

When space is noncompact the Hamiltonian of the theory can be used to define a suitable topology on $\text{Map}(\mathcal{S}_{d-1}, M_n)$ since, among other conditions (such as differentiability) we must restrict to finite energy field configurations, $\text{Map}_{f.e.}(\mathcal{S}_{d-1}, M_n)$.

For example, endow $\mathcal{S}_d = \mathbb{R} \times \mathcal{S}_{d-1}$ with a Minkowskian metric so that

$$S_{Mink}[\phi] = \int dt \int_{\mathcal{S}_{d-1}} \left[\frac{1}{2} G_{\mu\nu}(\phi) (\dot{\phi}^\mu \dot{\phi}^\nu - h^{\alpha\beta} \partial_\alpha \phi^\mu \partial_\beta \phi^\nu) - U(\phi) \right] \quad (18.3)$$

Now, the energy is found by computing $\pi_\mu = G_{\mu\nu}(\phi) \dot{\phi}^\nu$ and hence

$$E = \int_{\mathcal{S}_{d-1}} \text{vol} \left[\frac{1}{2} G^{\mu\nu}(\phi) \pi_\mu \pi_\nu + \frac{1}{2} G_{\mu\nu}(\phi) h^{\alpha\beta} \partial_\alpha \phi^\mu \partial_\beta \phi^\nu + U(\phi) \right] \quad (18.4)$$

If $G_{\mu\nu}$ is positive definite then all three terms are positive definite.

Let us now look at some examples with the noncompact space $\mathcal{S}_{d-1} = \mathbb{R}^{d-1}$, with the Euclidean metric.

We first study the case of $d = 2$ below. If $d = 2$ then the boundary at infinity can be considered to be the zero sphere S^0 , which is disconnected. Finiteness of the energy requires that ϕ tends to a zero of $U(\phi)$ for $x \rightarrow \pm\infty$. But this need not be the same zero! Field configurations which interpolate between different zeroes are in nontrivial soliton sectors. They cannot be connected by any finite energy process.

We will then consider examples with $d > 2$.

18.2 A simple motivating example: Solitons in the theory of a scalar field in 1 + 1 dimensions

As an explicit example of soliton sectors consider the action for a single real scalar field on 1 + 1-dimensional Minkowski space:

$$\begin{aligned} S[\phi] &= \int_{\mathbb{M}^{1,1}} \left(-\frac{1}{2} \partial_\mu \phi \partial^\mu \phi - U(\phi) \right) d^2x \\ &= \int_{\mathbb{M}^{1,1}} \left(\frac{1}{2} (\dot{\phi})^2 - \frac{1}{2} (\partial_x \phi)^2 - U(\phi) \right) d^2x \end{aligned} \quad (18.5)$$

where $U(\phi)$ is positive semi-definite and has zeroes ϕ_i where $U(\phi_i) = 0$.

As a concrete example, consider

$$U(\phi) = g(\phi^2 - v^2)^2 \quad (18.6)$$

Here g, v are real. It is important that $g > 0$ so the energy is bounded below. WLOG we can take $v > 0$. Here we are putting the speed of light $c = 1$ so time is measured in units of length. An action must have units of energy $\times$ time so if $c = 1$, an action has units $[ML]$. Thus, $[\phi] = [v] = (ML)^{1/2}$ while $[g] = (ML^3)^{-1}$.

Exercise

Show that expanding (18.6) around either vacuum $\phi = \pm v + \delta\phi$ gives a mass term with $m^2 = 8gv^2$ together with cubic and quartic interactions.

Remark: The “mass” m has dimensions of $1/L$. Upon quantization of the small fluctuations of ϕ around either of the vacua one would find particles with mass $m\hbar$. Since $\hbar$ has units of action ML (when we set $c = 1$) this has the correct units of mass.

In general, finiteness of the energy certainly implies that $\phi \rightarrow \phi_i$ for $x \rightarrow \pm\infty$, but if there are several zeroes of $U(\phi)$ then there are many choices for these boundary conditions. For example, in (18.6) we can have $\phi(x) \rightarrow \pm v$ at each end of the real line. Note that at each end of the line we have a *choice* of sign. Thus, if $\epsilon_L \in \{\pm 1\}$ and $\epsilon_R \in \{\pm 1\}$ we can choose boundary conditions (ϵ_L, ϵ_R) , meaning:

$$\lim_{x \rightarrow -\infty} \phi(x) = \epsilon_L v \quad \& \quad \lim_{x \rightarrow +\infty} \phi(x) = \epsilon_R v \quad (18.7)$$

We claim that, within the set of finite-energy field configurations the distinct choices of (ϵ_L, ϵ_R) define four different *disconnected components*. This is intuitively obvious: In order to change say from $\epsilon_R = +1$ to $\epsilon_R = -1$ we must continuously change the field away from $\phi^2 = v^2$ over an infinite volume space, namely, an infinite region $x \geq x_0$ for some x_0 , and such field configurations will have infinite energy barriers.

These four disconnected components of fieldspace are known as topological sectors.

The topological sector can be “measured” by a topological charge associated to a current. The conserved “topological current” is $j_\mu = \epsilon_{\mu\nu} \partial^\nu \phi$ and the conserved charge is

$$Q = \int_{-\infty}^{+\infty} j_0 = \int_{-\infty}^{+\infty} \partial_x \phi dx = \phi(+\infty) - \phi(-\infty) = (\epsilon_R - \epsilon_L)v \quad (18.8)$$

The Hamiltonian, or energy of a time-independent field configuration $\phi(x)$ is given by

$$H[\phi] = \int_{\mathbb{R}} \left[\frac{1}{2} (\partial_x \phi)^2 + U(\phi) \right] dx \quad (18.9)$$

Classical vacuum states are absolute minima of the energy. Clearly H is positive semidefinite so we can minimize it by putting $\partial_x \phi = 0$ and $U(\phi) = 0$. The first equation says that $\phi(x)$ is a constant in x and the second says that this constant value is $\phi = \phi_i$, one of the zeroes of $U(\phi)$. For the special case (18.6) we would have $\phi = \pm v$: There are two vacuum states.

However, when we choose boundary conditions so that Q is nonzero there is no finite energy process that can change the field configuration to these vacua. We can minimize the energy in these “topologically nontrivial sectors” to produce our soliton. The energy is minimized for

$$\partial_x^2 \phi = U'(\phi) \quad (18.10)$$

Multiplying by $\partial_x \phi$ this becomes

$$\frac{d}{dx} \left(\frac{1}{2} (\partial_x \phi)^2 - U(\phi) \right) = 0 \quad (18.11)$$

and so

$$\frac{1}{2} (\partial_x \phi)^2 - U(\phi) = \kappa \quad (18.12)$$

for a constant κ . The boundary conditions imply the constant κ is zero. So,

$$\frac{1}{2}(\partial_x \phi)^2 = U(\phi) \quad (18.13)$$

If $\epsilon_L = \epsilon_R$ we can take ϕ to be constant and minimize the energy to zero. If $\epsilon_L \neq \epsilon_R$ we cannot have $\partial_x \phi = 0$ for all x . In this case, to find the minimal energy field configuration we can integrate equation (18.13) (with $\kappa = 0$) to get

$$\int_{\phi_1}^{\phi_2} \frac{d\phi}{\sqrt{2U(\phi)}} = \pm(x_2 - x_1) \quad (18.14)$$

where $\phi(x_1) = \phi_1$ and $\phi(x_2) = \phi_2$. We can now impose the boundary conditions to obtain a solution.

For such a solution we can simplify the total energy to

$$E[\phi] = \int_{\phi_-}^{\phi_+} \sqrt{2U(\phi)} d\phi \quad (18.15)$$

where the sign is chosen so that $E[\phi] \geq 0$ and ϕ_- and ϕ_+ are two consecutive zeros of $U(\phi)$. Note that one does not need to know the detailed solution to compute the energy.

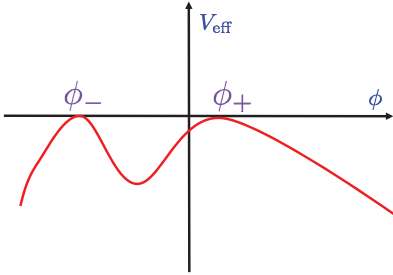


Figure 117: If $U(\phi)$ has more than one zero then we can show that soliton configurations from simple considerations of an effective one-dimensional particle mechanics problem with potential energy $-U(\phi)$.

Rather than directly integrating the solution equation the following nice argument, due to Sidney Coleman ⁸³ shows that a soliton solution must exist. We return to the Hamiltonian for time-independent field configurations (18.9). Let us interpret this as the *action* for a point particle. The spatial variable x of the field theory is reinterpreted as the “time coordinate” of the point particle of position ϕ . Note that the potential energy of this mechanics problem is $V_{\text{eff}} = -U(\phi)$. Thus the potential energy in the mechanics problem V_{eff} is bounded above and not below. We are interested in two successive zeroes as in Figure 117. Note that if the particle starts near one maximum of V_{eff} at ϕ_+ or ϕ_- then

⁸³S. Coleman, “Classical Lumps and their Quantum Descendents,” in *Aspects of Symmetry*.

if it has positive energy it will overshoot at the other maximum. If we start the particle near $\phi_{\pm}$ with negative energy it will not reach the other maximum. So there should be a solution in between where the particle reaches one maximum in the infinite “past” $x = -\infty$ and the other maximum in the infinite “future” $x = +\infty$.

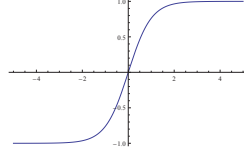


Figure 118: The soliton with boundary conditions $(+, -)$ for the special potential with $v = 1$, $m = 2$, (hence $g = 1/2$) and $x_0 = 0$.

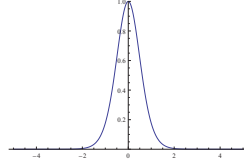


Figure 119: The energy of the soliton with boundary conditions $(\pm, \mp)$ for the special potential with $v = 1$, $m = 2$, (hence $g = 1/2$) and $x_0 = 0$.

The above equations apply to a general potential, and in general that is as far as you can go. For the special potential (18.6) we can do the integral explicitly to produce the explicit solution:

$$\phi = \pm v \tanh\left(\frac{m}{2}(x - x_0)\right) \quad (18.16)$$

with energy

$$E = \frac{1}{12} \frac{m^3}{g} \quad (18.17)$$

This special example nicely illustrates some aspects of solitons which turn out to hold quite generally:

1. The solution (18.16) transitions rapidly, on a scale set by the potential (here the scale $1/m$) from one vacuum to the other (here from $\pm v$ to $\mp v$).
2. Solutions to (18.13) will always come in families: If $\phi(x)$ is a solution then so is $\phi(x - x_0)$ for any x_0 , simply because the equation is translation invariant. This leads to a *moduli space of solutions*. Are there other solutions? In order to find other solutions we could vary the soliton equation around one solution, say ϕ_* , so we write $\phi = \phi_* + \delta\phi$ and expand to first order in $\delta\phi$. The result is a *linear* first order differential equation for $\delta\phi$

$$\partial_x(\delta\phi) - \frac{U'(\phi_*)}{\sqrt{2U(\phi_*)}}\delta\phi = 0 \quad (18.18)$$

with boundary conditions that $\lim_{x \rightarrow \pm\infty} \delta\phi(x) = 0$. Note that the coefficients in the linear equation depend on ϕ_* . This equation can be rewritten as

$$\frac{d(\delta\phi)}{\delta\phi} = \frac{U'(\phi_*)}{\sqrt{2U(\phi_*)}} dx = \frac{d}{dx}(\log \partial_x \phi_*) dx \quad (18.19)$$

and so the general solution is $\delta\phi(x) = c \frac{d}{dx} \phi_*(x)$ where c is a constant. We could have predicted this solution from the family obtained by shifting x_0 . The point of this exercise is that we now know there are no other moduli and so, in this example (assuming some genericity of U) the moduli space is just a copy of $\mathbb{R}$, parametrized by x_0 . In more complicated examples of solitons the analog of equation (18.18) will be a Dirac equation with an interesting space of solutions corresponding to nontrivial moduli spaces of solitons.

3. If we plot the energy *density* then it will be concentrated in the region, of size $\sim 1/m$, where ϕ jumps from one vacuum to another. In the case of (18.16) the energy density is

$$\mathcal{E} = \frac{m^4}{32g} \frac{1}{(\cosh \frac{1}{2}m(x - x_0))^4} \quad (18.20)$$

The solution is a lump of energy concentrated around x_0 . Such solutions of nonlinear differential equations are known as *solitons*. For a certain class of *integrable systems* they have remarkable properties. Many aspects of field theory, such as S-matrix amplitudes, become exactly soluble. See especially the book by Faddeev and Takhtadjan in Section §20 for more on this.

4. The energy goes to infinity for $g \rightarrow 0$. Under the correspondence principle the classical solution is related to (a family of) quantum states. As opposed to the particles with mass $8gv^2\hbar$ mentioned before, these states cannot be deduced from perturbation theory around a vacuum.
5. The energy of the corresponding quantum states can be computed as a series in perturbation theory: One substitutes the $\phi = \phi_*(x) + \phi_q(x, t)$ in to the Lagrangian and quantizes $\phi_q(x, t)$ to obtain an energy expansion

$$E = \frac{m^3}{12g} + \alpha_0 + \alpha_1 g + \alpha_2 g^2 + \dots \quad (18.21)$$

which is an asymptotic expansion for $g \rightarrow 0$.

Exercise Sine-Gordon model

Consider the sine-Gordon model, a 1 + 1-dimensional field theory with a field $\phi : \mathbb{M}^{1,1} \rightarrow \mathbb{R}$ with Hamiltonian

$$H = \int_{\mathbb{R}} dx \left(\frac{1}{2} \pi^2 + \frac{1}{2} (\partial_x \phi)^2 + \frac{m^2}{\beta^2} (1 - \cos(\beta\phi)) \right) \quad (18.22)$$

Here π is the momentum of the field. Under the Legendre transform to the Lagrangian formulation we have $\dot{\phi} = \pi$.

- a.) Find the classical vacua, that is, the field configurations with $H[\phi] = 0$.
- b.) Compute the topological charge between a pair of two such vacua.
- c.) Expand the potential $U(\phi)$ around a vacuum. (The quadratic term gives the mass of the perturbative particles.)
- d.) Find an exact formula for the minimum energy field configuration $\phi(x)$ interpolating between two consecutive vacua.
- e.) Compute the energy density of the field configuration of part (d).
- f.) Is there a time-independent solution with topological charge $4\pi/\beta$?
- g.) Show that the sine-Gordon equation of motion for a field $\phi(x, t)$ is equivalent to the “zero-curvature condition”:

$$[D_x, D_t] = 0 \quad (18.23)$$

where D_x and D_t are matrix-valued differential operators:

$$\begin{aligned} D_x &= \frac{\partial}{\partial x} + k \cos(\phi/2)\tau^1 + \omega \sin(\phi/2)\tau^2 + \frac{1}{2}(\partial_t \phi)\tau^3 \\ D_t &= \frac{\partial}{\partial t} + \omega \cos(\phi/2)\tau^1 + k \sin(\phi/2)\tau^2 + \frac{1}{2}(\partial_x \phi)\tau^3 \end{aligned} \quad (18.24)$$

Here $\omega^2 - k^2 = 1$, and $\tau^a = -\frac{i}{2}\sigma^a$ so that

$$[\tau^a, \tau^b] = \epsilon^{abc}\tau^c \quad (18.25)$$

Remark: This is an important observation and is the beginning of the application of integrable systems theory to the sine-Gordon model. See, for example, the book of Faddeev and Takhtadjan.

18.3 Landau-Ginzburg solitons in 1+1 dimensions

As a second example of 1 + 1 dimensional solitons we consider “Landau-Ginzburg models.” We will consider those related to models with extended supersymmetry. They will be field theories with N complex fields ϕ^i , $i = 1, \dots, N$ and a Lagrangian of a special type, described below.⁸⁴

The general LG action is based on a “Kähler metric” which for us will be a positive Hermitian form $K_{ij}(\phi, \bar{\phi})$ and a holomorphic function $W(\phi)$. The action on $\mathbb{M}^{1,1}$ is then:

$$- \int_{\mathbb{M}^{1,1}} \left(K_{ij} \partial_\mu \phi^i \partial^\mu \bar{\phi}^j + K^{ij} \frac{\partial W}{\partial \phi^i} \frac{\partial \bar{W}}{\partial \bar{\phi}^j} \right) d^2 x \quad (18.26)$$

The motivation for this action is that it is the bosonic part of a supersymmetric field theory. The supersymmetrization includes extra fermionic fields.

⁸⁴In general the models are defined for scalar fields $\phi : \mathbb{M}^{1,1} \rightarrow X$ where X is a complex manifold equipped with a metric known as a “Kähler metric.”

The classical vacua on a spatial slice $\mathbb{R}$ are the critical points $\phi_{(a)}$ of W . For example, if $N = 1$ and W is a polynomial of degree M then there are generally $M - 1$ isolated critical points: These are the roots of $W'(\phi) = 0$ and in general the roots will have no multiplicity so that $W''(\phi_{(a)}) \neq 0$. More generally, we will assume all critical points of W are (holomorphic) Morse critical points in the sense that $\partial_i \partial_j W(\phi_a)$ is an invertible matrix at every critical point ϕ_a . When expanding around such classical vacua the small fluctuations are massive.

The energy of a time-independent field configuration is

$$E = \int_{-\infty}^{+\infty} dx \left(|\partial_x \phi^i|^2 + \left| \frac{\partial W}{\partial \phi^i} \right|^2 \right) \quad (18.27)$$

Finiteness of the energy requires that we choose boundary conditions:

$$\lim_{x \rightarrow -\infty} \phi(x) = \phi_{(a)} \quad \lim_{x \rightarrow +\infty} \phi(x) = \phi_{(b)} \quad (18.28)$$

We can find the first order soliton equations using a nice trick, known as the *Bogomolnyi trick*.⁸⁵

For simplicity, in what follows we just take $K_{ij} = \delta_{ij}$. If ζ is any phase then we write the energy as

$$\begin{aligned} E &= \int_{-\infty}^{+\infty} dx \left(|\partial_x \phi^i|^2 + \left| \frac{\partial W}{\partial \phi^i} \right|^2 \right) \\ &= \int_{-\infty}^{+\infty} dx \left| \partial_x \phi^i - \zeta \frac{\partial \bar{W}}{\partial \bar{\phi}^i} \right|^2 + \int_{-\infty}^{+\infty} dx \left(\zeta \partial_x \bar{\phi}^i \frac{\partial \bar{W}}{\partial \bar{\phi}^i} + \zeta^{-1} \partial_x \phi^i \frac{\partial W}{\partial \phi^i} \right) \\ &= \int_{-\infty}^{+\infty} dx \left| \partial_x \phi^i - \zeta \frac{\partial \bar{W}}{\partial \bar{\phi}^i} \right|^2 + \int_{-\infty}^{+\infty} dx \frac{d}{dx} (\zeta \bar{W} + \zeta^{-1} W) \\ &= \int_{-\infty}^{+\infty} dx \left| \partial_x \phi^i - \zeta \frac{\partial \bar{W}}{\partial \bar{\phi}^i} \right|^2 + 2 \text{Re}[\zeta^{-1} (W_b - W_a)] \end{aligned} \quad (18.29)$$

where $W_a = W(\phi_{(a)})$.

As we have said, finiteness of the energy requires that $\phi(\pm\infty)$ is one of the critical points $\phi_{(a)}$. If there are different critical points at either end, say $b \neq a$ then ϕ is in a nontrivial soliton sector of type (ab) . The above applies for any phase ζ and the best lower bound on the energy comes from taking ζ to be the phase of $W(b) - W(a)$

So *if* we can construct a solution to

$$\partial_x \phi^i = \zeta \frac{\partial \bar{W}}{\partial \bar{\phi}^i} \quad (18.30)$$

satisfying the boundary conditions (18.28) then the energy is minimized and $E = 2|W(b) - W(a)|$.

⁸⁵In general the Bogomolnyi trick refers to a manipulation where the action, or the energy, is written as an integral of a perfect square up to boundary terms. The perfect square involves first order derivatives and the minimization of the energy is thus simplified enormously to a first order differential equation. Moreover, the boundary terms then typically give a nice formula for the energy.

This model can be supersymmetrized by adding fermionic partners ψ^i to the ϕ^i . Then it can be shown that this formula for the energy of the quantum state associated to the soliton is *exact*.

It remains to show that we can solve the differential equation. To discuss this let us consider the boundary value problem with general phase ζ .

We then observe a very striking aspect of the equation: Given any solution $\phi(x)$ on some open region of the x -axis we can say:

$$\frac{d}{dx}W(\phi(x)) = \frac{\partial W}{\partial \phi^i} \frac{d\phi^i}{dx} = \zeta \left| \frac{\partial W}{\partial \phi^i} \right|^2 \quad (18.31)$$

Therefore

$$W(x) - W(x_0) = \zeta \int_{x_0}^x \left| \frac{\partial W}{\partial \phi^i} \right|^2 dx' \quad (18.32)$$

and therefore *the image of the soliton in the W -plane is a straight line with phase ζ .*

Therefore, if $\phi(x) \rightarrow \phi_{(a)}$ for $x \rightarrow -\infty$ then the solution must project to a ray in the W -plane emerging from W_a with slope ζ . Similarly, if $\phi(x) \rightarrow \phi_{(b)}$ for $x \rightarrow +\infty$ it must project to a ray terminating on W_b . If W_a, W_b are in general position and ζ is a generic phase these rays will not intersect.

Thus, for general phase ζ there will be no solution to the boundary value problem. However, if ζ is the phase of the difference of the critical values $W_b - W_a$ then there might be a solution.

Let $\mathcal{L}_a^\zeta$ be the set of solutions $\phi^i(x)$ of the soliton equation (18.30) such that

$$\lim_{x \rightarrow -\infty} \phi^i(x) = \phi_a \quad (18.33)$$

Similarly, let $\mathcal{R}_b^\zeta$ be the set of solutions such that

$$\lim_{x \rightarrow +\infty} \phi^i(x) = \phi_b \quad (18.34)$$

We have a soliton for every point in $\mathcal{L}_a^\zeta \cap \mathcal{R}_b^\zeta$. Note that, thanks to the translation invariance, if $\phi(x) \in \mathcal{L}_a^\zeta \cap \mathcal{R}_b^\zeta$ then so is $\phi(x - x_0)$ for any x_0 . Thus, every point in $\mathcal{L}_a^\zeta \cap \mathcal{R}_b^\zeta$ in fact sits on a line.

Each solution $\phi \in \mathcal{L}_a^\zeta$ has an image $\{\phi(x) | x \in \mathbb{R}\} \subset X$, where $X = \mathbb{C}^N$ here (and is the Kähler target manifold, in general). The union of these images is a subspace of X known as a (left) *Lefschetz thimble* and denoted L_a^ζ . Similarly, the images of $\phi \in \mathcal{R}_b^\zeta$ is a right Lefschetz thimble and denoted R_b^ζ . Note that

1. The one-manifold $\{\phi(x) | x \in \mathbb{R}\} \subset X$ is the same for $\phi(x)$ and all its translates $\phi(x - x_0)$.

2. Every $\{\phi(x) | x \in \mathbb{R}\} \subset X$ for $\phi \in \mathcal{L}_a^\zeta$ projects to a ray in the complex plane parallel to ζ and emerging from the critical value W_a . Moreover, for a fixed solution we could equally well parametrize the one-manifold by x or by the value of W at $\phi(x)$.

3. With this parametrization $W \rightarrow W_a$ corresponds to $x \rightarrow -\infty$.

Now if there is a nonzero intersection $L_a^\zeta \cap R_b^\zeta$ then in fact they must intersect along a line in X . It begins at ϕ_a and terminates at ϕ_b . The image under W is a line segment

connecting W_a to W_b . Every such line in X corresponds to a solution of the soliton equation (18.30) with boundary conditions (18.28).

Therefore, let us get an idea of what L_a^ζ looks like. When W is near the critical value W_a we can replace W by its quadratic approximation. By a suitable redefinition of coordinates we can assume that the Hessian has been diagonalized so that we can write

$$W = W_a + \sum_i \frac{1}{2} \mu_i (\phi^i - \phi_a^i)^2 + \dots \quad (18.35)$$

The general solution of the equation for such a quadratic W is easily written down:

$$\phi^i = \phi_a^i + r^i \sqrt{\frac{\zeta \mu_i}{\kappa_i}} e^{\kappa_i(x-x_0)} \quad (18.36)$$

where $r^i \in \mathbb{R}$ and κ_i are real numbers with

$$|\kappa_i| = |\mu_i|. \quad (18.37)$$

Note that the $|\mu_i| \neq 0$ and hence we have the exponential approach to the critical point. Note that a shift of x_0 corresponds to a common rescaling of r^i .

If we want the solution to approach the critical point ϕ_a at $x \rightarrow -\infty$ then we choose the sign with all $\kappa_i > 0$. If we want the solution to approach the critical point ϕ_a at $x \rightarrow +\infty$ we choose the sign with all $\kappa_i < 0$.

If W_0 is a regular value near ϕ_a , say $W_0 = W_a + R\zeta$ where $R > 0$ is real then

$$\{\phi : W(\phi) = W_0\} \quad (18.38)$$

is isomorphic to T^*S^{n-1} . This is most easily seen by shifting and rescaling the ϕ^i so that $W = W_a + \sum (\phi^i)^2$. Then, decomposing $\phi^i = x^i + \sqrt{-1}y^i$ into real and imaginary parts the equations determining the inverse image of $W - W_a = R$ are

$$\begin{aligned} \vec{x}^2 - \vec{y}^2 &= R \\ \vec{x} \cdot \vec{y} &= 0 \end{aligned} \quad (18.39)$$

Solutions are uniquely parametrized by $\hat{x} \in S^{n-1}$, where S^{n-1} is the unit sphere and $\vec{y}$, which is in the tangent space to S^{n-1} at $\hat{x}$. Note that the equations uniquely fix $|\vec{x}|$, and it is natural to say that the zero-section of T^*S^{n-1} has radius-square R .

Now, in the inverse image above a regular point W_0 what is the intersection with L_a^ζ ? Again, with $\mu_i = 1$ this just means that $\phi^i = \phi_a^i + r^i e^{x-x_0}$. Therefore $\vec{y} = 0$, and hence $W_0 \cap L_a^\zeta$ is just the zero-section of T^*S^{n-1} .

Thus, the picture we have of L_a^ζ , at least near a critical point is that it is an expanding spherical “wave-front” projecting to a half-line in the W -plane. In the case $n = 2$ the picture is one of a thimble, and hence L_a^ζ are in general called “Lefschetz thimbles.”

NEED FIGURE HERE

Note that the fiber above a regular value is symplectic and the intersection with L_a^ζ is half-dimensional and Lagrangian. As W evolves along the ray the quadratic approximation

breaks down, but one can show that the fiber above W_0 continues to be a complex and symplectic manifold and the intersection with L_a^ζ is a maximal Lagrangian subspace.

Let us now consider the general superpotential W . At any point w in the complex W plane, the inverse image of w in ϕ -space is a complex affine variety:

$$X_w = \{\phi : W(\phi) = w\} \quad (18.40)$$

If w is a regular value then X_w is a smooth manifold. However, as w approaches a critical value W_a the variety X_w becomes singular. Our description above shows that some spherical cycle in X_w shrinks to a point. This spherical cycle is a nontrivial homology cycle known as a *vanishing cycle*. As W evolves along the ray the spherical cycle evolves to some maximal Lagrangian subvariety $\Delta_a \subset X_w$.

The Lagrangians obtained from the Lefschetz thimbles L_a^ζ and R_a^ζ are two maximal Lagrangian subvarieties in X_w , when w is a regular value. These two half-dimensional subspaces generically intersect in points. There is a soliton solution for each such intersection. From this construction we learn:

1. We generically expect there to be solutions.
2. If the oriented intersection number of L_a^ζ and R_a^ζ in the fiber above a regular value is nonzer, then there are definitely solutions.

It turns out that some of these solutions continue to define special quantum states in the quantum theory for which one can make many beautiful *exact* statements. For example, the exact mass of the state is $2|W_a - W_b|$, and there are no quantum corrections.

This and many other beautiful things can be found in

1. S. Cecotti and C. Vafa, “On the Classification of N=2 Supersymmetric Theories” arXiv:hep-th/9211097
2. K. Hori, A. Iqbal and C. Vafa, “D-branes and mirror symmetry,” hep-th/0005247.
3. K. Hori, S. Katz, A. Klemm, R. Pandharipande, R. Thomas, C. Vafa, R. Vakil and E. Zaslow, “Mirror symmetry,” (Clay mathematics monographs. 1)
4. D. Gaiotto, G. Moore, and E. Witten, “Algebra of the Infrared: String Field Theoretic Structures in Massive $\mathcal{N} = (2, 2)$ Field Theory In Two Dimensions,” to appear.

18.4 Minkowskian spacetime of dimension greater than two

In this case the boundary at spatial infinity, S^{d-2} is connected.

Let us consider finiteness of the energy. We look at static solutions $\pi_\mu = 0$ and then we have the same story as for the Euclidean action analyzed above, but one dimension lower. So, for $d > 2$ we must have:

- a.) $\lim_{r \rightarrow \infty} \phi(t, r\hat{x}) = \phi_0$ for some constant ϕ_0 , independent of angles.
- b.) $U(\phi_0) = 0$.

so, we can now define a field configuration $\bar{\phi}$ on $S^{d-1} \rightarrow M_n$, and the disconnected components of configuration space for these boundary conditions are classified by $\pi_{d-1}(M_n, \phi_0)$.

Example: *Skyrmions*. In nuclear and particle physics mesons are described by a 3+1-dimensional sigma model with target manifold $(SU(2) \times SU(2))/SU(2) \cong SU(2)$ and $(SU(3) \times SU(3))/SU(3) \cong SU(3)$. Generalizing to $SU(N_c)$ QCD with N_f flavors of quarks

we have $(SU(N_f) \times SU(N_f))/SU(N_f) \cong SU(N_f)$. As we shall prove below, for $N \geq 2$ we have $\pi_3(SU(N)) = \mathbb{Z}$. The solitons in the nontrivial sectors are called *Skyrmions*.

See E. Witten, “Global aspects of current algebra,” Nucl. Phys. **B223** (1983) 422 for a beautiful discussion of these solitons.

18.5 Solitons in spontaneously broken gauge theories

Finally, in gauge theories there is an important modification to the above formula for the energy: Derivatives are replaced by gauge covariant derivatives and the contribution to the total energy from the scalar fields is a sum of three positive definite terms:

$$E = \int_{S_{d-1}} \text{vol} \cdot \left[\frac{1}{2} G^{\mu\nu}(\phi) \pi_\mu \pi_\nu + \frac{1}{2} G_{\mu\nu} D_\alpha \phi^\mu D_\alpha \phi^\nu + U(\phi) \right] + E_{Y.M.} \quad (18.41)$$

where

$$(D_\alpha \phi)^\mu = \partial_\alpha \phi^\mu + A_\alpha^a K_a^\mu(\phi) \quad (18.42)$$

and $K_a^\mu(\phi)$ is a Killing vector for the gauge symmetry action on M_n (and a runs over a basis for the Lie algebra). Finally $E_{Y.M.}$ is the energy of the Yang-Mills fields.

Take $\mathcal{S}_{d-1} = \mathbb{R}^{d-1}$ with Euclidean metric. Define the set of classical vacua of the scalar field theory by:

$$\mathcal{M}_{cl.vac} := \{\phi | U(\phi) = 0\} \subset M_n \quad (18.43)$$

Suppose that the target space $M_n = V$ is a linear representation of the gauge group G and $U(\phi)$ is a G -invariant potential then the critical points of $U(\phi)$ will admit a G -action.

Suppose ϕ_0 is a critical point and the stabilizer of ϕ_0 is a subgroup $H \subset G$. We say that the gauge symmetry has been spontaneously broken from G to H . In this case there is an orbit through ϕ_0 which can be identified with the homogeneous space G/H . In some important cases the entire critical locus is this orbit. In that case $\pi_{d-2}(G/H)$ classifies the soliton sectors. In particular, in four-dimensional Grand Unified Theories, the presence of monopoles often comes down to a computation of $\pi_2(G/H)$. We will see in Chapter *** how to compute such homotopy groups.

Finiteness of the energy on $\mathbb{R}^{d-1}$ requires that the values of ϕ at $\vec{x} \rightarrow \infty$ live in $\mathcal{M}_{cl.vac}$. However, it can happen that there are finite energy gauge fields such that $D_\alpha \phi \rightarrow 0$ as $r \rightarrow \infty$ and moreover we have that

$$\lim_{r \rightarrow \infty} \phi(t, r\hat{x}) := \phi_\infty(t, \hat{x}) \quad (18.44)$$

is a nontrivial map

$$S^{d-2} \rightarrow \mathcal{M}_{cl.vac} \quad (18.45)$$

The presence of suitable gauge fields is crucial here: The two terms in the covariant derivative balance each other.

Since time evolution must be continuous it follows that the map

$$\phi_\infty(t, \cdot) : S_\infty^{d-2} \rightarrow \mathcal{M}_{cl.vac} \quad (18.46)$$

must be in a fixed homotopy class

$$\pi_{d-2}(\mathcal{M}_{cl.vac.}) \quad (18.47)$$

If this class is nontrivial the soliton is topologically stable: It would take an infinite amount of energy to change the homotopy class at spatial infinity.

Example 1: Vortices An important example is of spontaneous symmetry breaking of $U(1)$. We can take a $U(1)$ gauge field A and a scalar field ϕ of charge 1. The Lagrangian is

$$-\int \frac{1}{2e^2} F * F + (D\phi)^* * D\phi + \lambda (|\phi|^2 - v^2)^2 \quad (18.48)$$

where $D\phi = (d + A)\phi$, and $F = dA$. We can view this as a Lagrangian in $3 + 1$ dimensions - in which case it is an effective theory for a superconductor (with ϕ representing the Cooper pair condensate. It should then have charge 2, not 1), or we can view this as a Lagrangian in $1 + 1$ dimensions, in which case it is known as the *Abelian Higgs model*. For $v^2 \neq 0$ the model breaks $U(1)$ to nothing.⁸⁶

We can look for two-dimensional instantons or, equivalently, $3+1$ dimensional string-like solitons. The latter are time independent and translationally invariant along - say - the x^3 axis. The problems are then formally equivalent.

Finiteness of the action of the instanton, or tension of the string, requires that F be pure gauge at infinity and $|\phi| \rightarrow v^2$ at infinity. Nevertheless, the phase of ϕ , which is well-defined at infinity, defines a winding number $S_\infty^1 \rightarrow S^1$.

If we choose the special value $\lambda = e^2/2$ then we can be more explicit. The 2d action/4d string tension is

$$\int dx^1 dx^2 \left(\frac{1}{2e^2} B_3^2 + |D_1\phi|^2 + |D_2\phi|^2 + \frac{e^2}{2} (|\phi|^2 - v^2)^2 \right) \quad (18.49)$$

When we discuss connections we will take A to be (locally) an imaginary one-form so $F_{12} = \partial_1 A_2 - \partial_2 A_1 = iB_3$. Now (18.49) can be nicely rewritten as

$$\int dx^1 dx^2 \left(\frac{1}{2e^2} (B_3 \mp e^2 (|\phi|^2 - v^2))^2 + |D_1\phi \pm iD_2\phi|^2 \right) \mp \frac{1}{2} v^2 \Phi \quad (18.50)$$

where $\Phi = \int dx^1 dx^2 B_3$ is the total flux. The soliton which minimizes the energy is then given by first order equations:

$$\begin{aligned} B_3 &= \pm e^2 (|\phi|^2 - v^2) \\ D_1\phi \pm iD_2\phi &= 0 \end{aligned} \quad (18.51)$$

with $E = \frac{v^2}{2} |\Phi|$. We choose the sign in (18.51) so that we get the absolute value. The equations (18.51) are known as the *vortex equations*. Note that from the second equation we learn that as long as $\phi \neq 0$ we can write (choosing the $-$ sign for definiteness):

⁸⁶If ϕ has charge q then in fact the gauge group is broken to $\mathbb{Z}/q\mathbb{Z}$.

$$A_z = -\phi^{-1}\partial_z\phi \quad (18.52)$$

and at infinity, where the norm of ϕ becomes a constant, this means that A is pure gauge. The topological quantum number is the winding number, which can be identified with the flux, by Stokes' theorem.

Example 2: *'t Hooft-Polyakov monopole.* This is based on a Yang-Mills-Higgs gauge theory in $d = 3 + 1$ dimensions with gauge group $G = SO(3)$ and a Higgs scalar field $\vec{\phi}$ in the triplet representation so $M_n = \mathbb{R}^3$.

Now

$$U(\phi) = g(\vec{\phi}^2 - m^2)^2 \quad (18.53)$$

so that $\mathcal{M}_{cl.vac} = S^2$. The soliton sectors are classified by $\pi_2(S^2) = \mathbb{Z}$.

18.6 The general field theory of scalar fields: The nonlinear sigma model

One important class of field theories is based on the space of fields which are maps from spacetime $\mathcal{S}_d$ to some “target space manifold” M_n :

$$\phi : \mathcal{S}_d \rightarrow M_n \quad (18.54)$$

This is a generalization of the usual scalar field theories where $\mathcal{S}_d = \mathbb{R}^d$, is Minkowski space or Euclidean space and $M_n = \mathbb{R}^n$ for a theory of n real scalar fields.

We denote the space of fields (18.54) by $Map(\mathcal{S}_d, M_n)$. We have discussed its topology above. In fact, it can be made into an infinite-dimensional manifold. In Chapter 4 below we will show that its tangent space can be understood using the language of bundles:

$$T_{\phi_0}Map(\mathcal{S}_d, M_n) = \Gamma(\phi_0^*(TM_n)) \quad (18.55)$$

For now we simply choose local coordinates ξ on $\mathcal{S}_d$ and coordinates x^μ on M_n in the neighborhood of $\phi_0(\xi)$. Then the tangent vector, or infinitesimal field variation is a map:

$$\xi \mapsto \sum_{\mu=1}^n \delta\phi^\mu(\xi) \frac{\partial}{\partial x^\mu} \Big|_{\phi_0(\xi)} \quad (18.56)$$

In order to write an action for the nonlinear sigma model we need to introduce *metrics* on both spacetime $\mathcal{S}_d$ and on the target space M_n . These metrics induce a metric on the infinite-dimensional manifold of fields $Map(\mathcal{S}_d, M_n)$. This follows directly from (18.55). We can write the metric in local coordinates as follows:

In local coordinates on $\mathcal{S}_d$ we can write

$$ds^2(\mathcal{S}_d) = h_{ij}(\xi) d\xi^i d\xi^j \quad (18.57)$$

with $1 \leq i, j \leq d$, and using local coordinates

$$ds^2(M_n) = G_{\mu\nu}(x) dx^\mu dx^\nu \quad (18.58)$$

on M_n . Then, a tangent vector $\delta\phi$ at ϕ_0 has norm:

$$\| \delta\phi \|^2 = \int_{\mathcal{S}_d} \text{vol}(h) \delta\phi^\mu(\xi) \delta\phi^\nu(\xi) G_{\mu\nu}(\phi_0(\xi)) \quad (18.59)$$

So, $\text{Map}(\mathcal{S}_d, M_n)$ is at least formally a Riemannian manifold - we can therefore speak of “distance in fieldspace.” This is a very important concept: Qualitatively different physical phenomena occur at finite and at infinite distance in fieldspace.

Now, the natural line element (18.59) on $\text{Map}(\mathcal{S}_d, M_n)$ gives an action principle for the nonlinear sigma model. The derivative of ϕ is a linear map

$$d\phi : T_\xi \mathcal{S}_d \rightarrow T_{\phi(\xi)} M_n \quad (18.60)$$

Since we have metrics on the vector spaces $T_\xi \mathcal{S}_d$ and on $T_x M_n$ we have a metric on $\text{Hom}(T_\xi \mathcal{S}_d, T_{\phi(\xi)} M_n)$.

Then the norm of (18.60) is

$$\| d\phi \|^2 = h^{ij}(\xi) G_{\mu\nu}(\phi(\xi)) \partial_i \phi^\mu \partial_j \phi^\nu \quad (18.61)$$

Now we have

$$\begin{aligned} S &= f_\pi^2 \int_{\mathcal{S}_d} \text{vol}(h) \| d\phi \|^2 \\ &= f_\pi^2 \int_{\mathcal{S}_d} d^d \xi \sqrt{|\text{deth}|} h^{ij}(\xi) G_{\mu\nu}(\phi(\xi)) \partial_i \phi^\mu \partial_j \phi^\nu \end{aligned} \quad (18.62)$$

where f_π is a dimensionful constant (see comments below).

The kinetic term of the nonlinear sigma model is equivalent to the line element on field space, the space of maps $\mathcal{S}_d \rightarrow M_n$.

Remarks:

- We can, of course, as a potential term $\int_{\mathcal{S}_d} d^d \xi \sqrt{|\text{deth}|} U(\phi)$ to the theory.
- The action must be dimensionless (in units with $\hbar = c = 1$. Since $G_{\mu\nu}$ is nontrivial in general the scalar fields $\phi^\mu(\xi)$ are dimensionless. Then f_π must have (spacetime) dimensions of $L^{1-d/2}$. When $G_{\mu\nu} = \delta_{\mu\nu}$ the theory is free, and f_π can be absorbed into the fields to give canonically normalized fields, but in general we have a nontrivially interacting theory. Thus, in the interacting case, except for $d = 2$, the classical sigma model has a dimensionful parameter. In general, these models are not renormalizable, and should be thought of as effective low energy field theories whose domain of validity is set by the scale f_π . For example, in theories of pions in $3 + 1$ dimensions, f_π has dimensions of mass, and is called the “pion decay constant.” In the theory of pions $M_n = SU(2) \times SU(2)/SU(2)_{\text{diag}}$ and $f_\pi \sim 93 \text{ MeV}$.

•

It often happens that in a physical problem there is also the data of a fiber bundle $E \rightarrow M_n$. This happens when there are fermions in the sigma model, or when we gauge symmetries. In such situations the fields of the theory often involve sections of the pulled back bundle $\phi^*(E)$. Note that such sections can indeed be thought of as fields in a fieldtheory on $\mathcal{S}_d$.

- An extremely important special case occurs when $d = 2$. Then the theory classically has no scale. Equivalently, it is conformally invariant under Weyl transformations $h \rightarrow \Omega^2 h$. However, for most target space metrics, $G_{\mu\nu}$ the theory is not quantum mechanically scale invariant. Indeed, the famous result ⁸⁷ of Friedan says that the dependence on logarithmic scale t is

$$\frac{d}{dt}G_{\mu\nu} = \mathcal{R}_{\mu\nu} + \dots \quad (18.63)$$

where $\mathcal{R}_{\mu\nu}$ is the Ricci tensor.

Exercise

What is the classical equation of motion for the nonlinear sigma model?

19. “Instanton” sectors

Suppose we have a Euclidean d -dimensional spacetime $\mathcal{S}_d$ and a nonlinear sigma model field:

$$\phi : \mathcal{S}_d \rightarrow M_n \quad (19.1)$$

The “instanton sectors” refers to the classification of the space of fields by its components. In the path integral

$$\int d\phi e^{-S[\phi]} \quad (19.2)$$

we “integrate over all field configurations.” This integration always involves a sum over the components of $Map(\mathcal{S}_d, M_n)$, and sometimes the path integral is very different in different components. This raises the issue of how to weight the different components.

We illustrate this important point with two elementary, but extended examples:

19.1 Fieldspace topology from boundary conditions

Even when spacetime is topologically trivial, say $\mathcal{S}_d = \mathbb{R}^d$ the *action* of the theory can effectively introduce topology, because the space of *finite action field configurations* $Map_{f.a.}(\mathbb{R}^d, M_n)$ can have nontrivial topology.

For example, suppose the action is

$$S[\phi] = \int_{\mathbb{R}^d} d^d \vec{x} \left[G_{\mu\nu}(\phi) \partial_\alpha \phi^\mu \partial_\alpha \phi^\nu + U(\phi) \right] \quad (19.3)$$

⁸⁷D. Friedan, “Nonlinear Models in Two Epsilon Dimensions,” Phys. Rev. Lett. **45**, 1057 (1980); D. H. Friedan, “Nonlinear Models in Two + Epsilon Dimensions,” Annals Phys. **163**, 318 (1985).

with Euclidean signature on spacetime $\mathbb{R}^d$ and on the target.

Let us study the behavior of ϕ as $\vec{x} \rightarrow \infty$. We define this by setting $\vec{x} = r\hat{x}$ and

$$ds^2 = dr^2 + r^2 g_{ij} d\theta^i d\theta^j \quad (19.4)$$

so that

$$S[\phi] = \int_{\mathbb{R}^d} \frac{dr}{r} \wedge \text{vol}(S^{d-1}) \left[r^d G_{\mu\nu}(\phi) \partial_r \phi^\mu \partial_r \phi^\nu + r^{d-2} g^{ij} G_{\mu\nu}(\phi) \partial_i \phi^\mu \partial_j \phi^\nu + r^d U(\phi) \right] \quad (19.5)$$

This is a sum of three positive semidefinite terms. The integral over the angles is a compact integral and poses no problem of convergence (for smooth field configurations). The problematic integral is over r : We therefore have convergence conditions for $r \rightarrow \infty$:

1. $\partial_r \phi \sim \frac{1}{r^{d/2+\epsilon}}$
2. $\partial_i \phi \sim \frac{1}{r^{(d-2)/2+\epsilon}}$
3. $U(\phi(r\hat{x})) \sim \frac{1}{r^{d/2+\epsilon}}$

Here ϵ is any positive quantity.

It follows from (1) and (2) that $\lim_{r \rightarrow \infty} \phi(r\hat{x}) = \phi_0$ must go to a well-defined constant value, independent of angles. It then follows from (3) that that constant must be such that $U(\phi_0) = 0$.

Thus we see that any smooth field ϕ of finite action defines a map

$$\bar{\phi} : (S^d, \infty) \rightarrow (M_n, \phi_0) \quad (19.6)$$

and components of the space of maps will be (at least in part) classified by $\pi_d(M_n, \phi_0)$.

19.2 A charged particle on a ring around a solenoid, at finite temperature

Consider a particle of mass m confined to a ring of radius r in the xy plane. The position of the particle is described by an angle ϕ , so we identify $\phi \sim \phi + 2\pi$, and the action is

$$S = \int \frac{1}{2} m r^2 \dot{\phi}^2 = \int \frac{1}{2} I \dot{\phi}^2 \quad (19.7)$$

with $I = m r^2$ the moment of inertia.

Let us also suppose that our particle has electric charge e and that the ring is threaded by a solenoid with magnetic field B , so the particle moves in a zero B field, but there is a nonzero gauge potential

$$A = \frac{B}{2\pi} d\phi \quad (19.8)$$

The action is therefore:

$$\begin{aligned} S &= \int \frac{1}{2} I \dot{\phi}^2 dt + \oint e A \\ &= \int \frac{1}{2} I \dot{\phi}^2 dt + \frac{eB}{2\pi} \oint \dot{\phi} dt \end{aligned} \quad (19.9)$$

The second term is an example of a Chern-Simons term.⁸⁸

Now we compute the (Euclidean time) propagator

$$Z(\phi_2, \phi_1 | T) := \langle \phi_2 | e^{-\frac{TH}{\hbar}} | \phi_1 \rangle \quad (19.10)$$

for the particle, where H is the Hamiltonian.

We will compute (19.10) by both Hamiltonian and path integral techniques. The comparison is interesting.

We begin with the Hamiltonian viewpoint. The conjugate momentum is

$$L = I\dot{\phi} + \frac{eB}{2\pi} \quad (19.11)$$

We denote it by L because it can be thought of as angular momentum.

Note that the coupling to the flat gauge field has altered the usual relation of angular momentum and velocity. Now we obtain the Hamiltonian from the Legendre transform:

$$\int L\dot{\phi}dt - S = \int \frac{1}{2I} \left(L - \frac{eB}{2\pi}\right)^2 dt \quad (19.12)$$

Upon quantization $L \rightarrow -i\hbar \frac{\partial}{\partial \phi}$, so the eigenfunctions of the Hamiltonian are just

$$\frac{1}{\sqrt{2\pi}} e^{im\phi} \quad m \in \mathbb{Z} \quad (19.13)$$

They give singly degenerate energy states with energy

$$E_m = \frac{\hbar^2}{2I} (m - \mathcal{B})^2 \quad (19.14)$$

where $\mathcal{B} := \frac{eB}{2\pi\hbar}$.

Remarks:

- The action (19.9) makes good sense for ϕ valued in the real line or for $\phi \sim \phi + 2\pi$, valued in the circle. Making this choice is important in the choice of what theory we are describing. Where - in the above analysis did we make the choice that the target space is the circle?

- Although the Chern-Simons term is a total derivative it has a nontrivial effect on the quantum physics as we can see since B has shifted the spectrum of the Hamiltonian.

- The total spectrum is *periodic* in $\mathcal{B}$, and shifting $\mathcal{B} \rightarrow \mathcal{B}+1$ is equivalent to $m \rightarrow m+1$.

Now it is straightforward to compute:

$$Z(\phi_2, \phi_1 | T) = \frac{1}{2\pi} \sum_{m \in \mathbb{Z}} e^{-\frac{TH}{2I} (m - \mathcal{B})^2 + im(\phi_2 - \phi_1)} \quad (19.15)$$

⁸⁸This problem turns out to be closely related to quantum dots. See Yoshimasa Murayama, *Mesoscopic Systems*, Section 10.10

Again, note periodicity in $\mathcal{B}$.

The function appearing in (19.15) is known as a *theta function*, more on this below. Note that the large T asymptotics are easily read off:

$$Z(\phi_2, \phi_1|T) \stackrel{T \rightarrow \infty}{\sim} \begin{cases} \frac{1}{2\pi} e^{-T\hbar\{\mathcal{B}\}^2/2I+i[\mathcal{B}](\phi_2-\phi_1)}(1+\mathcal{O}(e^{-\kappa T})) & 0 \leq \{\mathcal{B}\} \leq \frac{1}{2} \\ \frac{1}{2\pi} e^{-T\hbar(1-\{\mathcal{B}\})^2/2I+i([\mathcal{B}]+1)(\phi_2-\phi_1)}(1+\mathcal{O}(e^{-\kappa T})) & \frac{1}{2} \leq \{\mathcal{B}\} \leq 1 \end{cases} \quad (19.16)$$

where $[\mathcal{B}]$ is the greatest integer in $\mathcal{B}$, and $\{\mathcal{B}\}$ is the fractional part.

Note too that the partition function is

$$Z := \text{Tre}^{-\beta H/\hbar} = \sum_{m \in \mathbb{Z}} e^{-\frac{\beta\hbar}{2I}(m-\mathcal{B})^2} \quad (19.17)$$

as we obtain by setting $Z = \int_0^{2\pi} d\phi Z(\phi, \phi|T)$.

Now let us compare the path integral derivation. In the rotation to Euclidean space the “Chern-Simons term” $\oint eA$ remains real so the Euclidean path integral is

$$Z(\phi_2, \phi_1|T) = \int [d\phi(t)]_{\phi(0)=\phi_1}^{\phi(T)=\phi_2} e^{-\frac{1}{\hbar} \int_0^T \frac{1}{2} I \dot{\phi}^2 + i \int \mathcal{B} \dot{\phi} dt} \quad (19.18)$$

We claim this is the same function $Z(\phi_2, \phi_1|T)$ that we defined in the Hamiltonian formulation. Note that the partition function (19.17) is the Euclidean path integral on a circle of radius $T = \beta/\hbar$.

This path integral is Gaussian so it can be done exactly by semiclassical techniques.

The fieldspace, $\text{Map}([0, T], S^1)$, is topologically nontrivial. Consider two such maps. Together they can be combined to form a map $S^1 \rightarrow S^1$, and this map has a homotopy class. Thus

$$\pi_0(\text{Map}([0, T], S^1)) \cong \pi_1(S^1) \cong \mathbb{Z}. \quad (19.19)$$

So there are infinitely many components which, noncanonically, can be labeled by a winding number w .

The equations of motion are simply $\ddot{\phi} = 0$. Note that the Chern-Simons terms has not changed them. There is one solution of the classical equations with boundary condition $\phi(0) = \phi_1, \phi(T) = \phi_2$ in each component of fieldspace. We can write it explicitly as follows:

We lift the map $[0, T] \rightarrow S^1$ to a covering map $\tilde{\phi} : [0, T] \rightarrow \mathbb{R}$. (Regarding $S^1 = \mathbb{R}/2\pi\mathbb{Z}$.) Then we write

$$\tilde{\phi}_c = \tilde{\phi}_1 + \left(\frac{\tilde{\phi}_2 - \tilde{\phi}_1 + 2\pi w}{T} \right) t \quad w \in \mathbb{Z} \quad (19.20)$$

(where $\tilde{\phi}_i$ are lifts of ϕ_i). These are solutions of the Euclidean equations of motion, and are known as *instantons*.

We now write

$$\phi = \phi_c + \phi_q \quad (19.21)$$

where ϕ_c is an instanton solution, as in (19.20), and ϕ_q is the quantum fluctuation with $\phi_q(0) = \phi_q(T) = 0$. Using this decomposition the path integral in each component of fieldspace factorizes into the contribution of the classical solution times the path integral over the quantum fluctuations ϕ_q .

Summing over all the components of fieldspace, weighted by the action principle we therefore get:

$$Z(\phi_2, \phi_1|T) = Z_q \sum_{w \in \mathbb{Z}} e^{-\frac{4\pi^2 I}{2\hbar T} (w + \frac{\phi_2 - \phi_1}{2\pi})^2 + 2\pi i \mathcal{B}(w + \frac{\phi_2 - \phi_1}{2\pi})} \quad (19.22)$$

Z_q is the path integral over ϕ_q and it is independent of topological sector. We can determine it in two ways:

First, let us study the $T \rightarrow 0$ behavior of the path integral. If ϕ_1 is near ϕ_2 then

$$Z \rightarrow Z_q e^{-\frac{I}{2\hbar T} (\phi_2 - \phi_1)^2 + i\mathcal{B}(\phi_2 - \phi_1)} + \mathcal{O}(e^{-\text{const}/T}) \quad (19.23)$$

i.e. the instantons are only important at *large* T . At small values of T the particle cannot “feel” the global topology of the target space. Put more precisely, an observer on the particle cannot tell, in a short period of time that the particle is confined to a circle and not a line because in a short amount of time the particle cannot move very far! Therefore, we can compare (19.23) with the standard quantum mechanical propagator. Locally we can remove the phase from the B -field via $\psi(\phi) \rightarrow e^{-i\mathcal{B}\phi}\psi(\phi)$. After this transformation we expect to recover the standard propagator of a particle of mass $M = I$ on the line. Rotated to Euclidean space this would be:

$$\sqrt{\frac{M}{2\pi\hbar T}} e^{-\frac{M(\phi_2 - \phi_1)^2}{2\hbar T}} \quad (19.24)$$

so

$$Z_q = \sqrt{\frac{I}{2\pi\hbar T}} \quad (19.25)$$

We can also recover this directly using ζ -function regularization of the determinant. See the exercise below.

Let us now compare the two expressions. The path integral version (19.22) converges rapidly for short times T while the Hamiltonian version (19.15) converges rapidly for long times T . To compare the two we introduce the standard Riemann theta function

$$\vartheta\left[\begin{smallmatrix} \theta \\ \phi \end{smallmatrix}\right](z|\tau) := \sum_{n \in \mathbb{Z}} e^{i\pi\tau(n+\theta)^2 + 2\pi i(n+\theta)(z+\phi)} \quad (19.26)$$

Using the Poisson summation formula we prove

$$\vartheta\left[\begin{smallmatrix} \theta \\ \phi \end{smallmatrix}\right]\left(\frac{-z}{\tau} \middle| \frac{-1}{\tau}\right) = (-i\tau)^{1/2} e^{2\pi i\theta\phi} e^{i\pi z^2/\tau} \vartheta\left[\begin{smallmatrix} -\phi \\ \theta \end{smallmatrix}\right](z|\tau) \quad (19.27)$$

The path integral answer (19.22) for $Z(\phi_2, \phi_1|T)$ can be written in terms of the standard theta function as

$$Z(\phi_2, \phi_1|T) = Z_q \vartheta\left[\frac{\phi_2 - \phi_1}{2\pi}\right](0|\tau) \quad (19.28)$$

with $\tau = i\frac{2\pi I}{\hbar T}$. Now, applying the transformation (19.27) with $z = 0$ we recover the expression (19.15) for $Z(\phi_2, \phi_1|T)$ derived using the Hamiltonian formalism.

This comparison has taught us several lessons:

Remarks

- We must sum over all instanton sectors to arrive at the correct result derived from the Hamiltonian formulation.
- The Euclidean path integral has a well-defined normalization, contrary to claims sometimes made in textbooks.
- Thus, the instanton number w and the angular momentum m are Fourier/Poisson dual.

Exercise *Groundstates*

- It is a good exercise to work out the ground-state energy as a function of $\mathcal{B}$. Plot the groundstate energy.
- Which value of m gives the ground state.

Exercise *ζ -function computation of the determinant*

Scaling the T -dependence out of the action we have the path integral measure:

$$\int [d\phi_q] \exp\left[-\int_0^1 \phi_q \left(-\frac{I}{2\hbar T} \frac{d^2}{d\tau^2}\right) \phi_q\right] = (2\pi) \text{Det}'^{-1/2}\left(-\frac{I}{2\hbar T} \frac{d^2}{d\tau^2}\right) \quad (19.29)$$

In general if an Hermitian operator $\mathcal{O}$ has eigenvalues λ_n , all of which are assumed to be nonzero, then we can define a ζ -function determinant by generalizing the finite dimensional expression

$$\det \mathcal{O} = \exp\left[\sum_n \log \lambda_n\right] = \exp\left[-\frac{d}{ds}\Big|_{s=0} \sum_n \lambda_n^{-s}\right] \quad (19.30)$$

We define the ζ -function $\zeta_{\mathcal{O}}(s) := \sum_n \lambda_n^{-s}$ which - you have to show! - converges to an analytic function of s for large real s and has an analytic continuation to $s = 0$. Then we *define*

$$\det \mathcal{O} := \exp[-\zeta'_{\mathcal{O}}(0)] \quad (19.31)$$

For $\mathcal{O} = -\frac{I}{2\hbar T} \frac{d^2}{d\tau^2}$ we have $\zeta_{\mathcal{O}}(s) = 2\left(\frac{I\pi^2}{2\hbar T}\right)^s \zeta(2s)$ and since

$$\zeta(s) = -\frac{1}{2} + \text{slog}\left(\frac{1}{\sqrt{2\pi}}\right) + \mathcal{O}(s^2) \quad (19.32)$$

we have

$$\text{Det}'(\mathcal{O}) := \exp[-\zeta'_{\mathcal{O}}(0)] = \frac{1}{2\pi^4} \frac{\hbar T}{I} \quad (19.33)$$

in agreement with (19.25) up to factors of (2π) .

19.3 Worldsheet instantons in string theory

In the theory of closed oriented strings one takes the worldsheet to be a closed oriented Riemann surface Σ . The target space M_n can be, for example $\mathbb{R}^4 \times K$ where K is a Calabi-Yau 3-fold. Then the sectors are classified by $[\Sigma, K]$. For example, if $\Sigma = S^2$, they are classified by $\pi_2(K)$.

In string theory the target space is often endowed with a (locally defined) 2-form B and then the 2-dimensional sigma models based on maps $\phi : \Sigma \rightarrow X$ has action:

$$S[\phi] = \frac{1}{4\pi\ell^2} \int_{\Sigma} \|d\phi\|^2 + 2\pi i \int_{\Sigma} f^*(B) \quad (19.34)$$

where ℓ is a constant with dimensions of (spacetime) length, called the *string length*.

In many important models B is a globally defined closed two-form. Then if Σ is closed, $\int_{\Sigma} \phi^*(B)$ only depends on the DeRham cohomology class of B and the map $\phi^* : H^2(X) \rightarrow H^2(\Sigma) \cong \mathbb{Z}$. We can then write $[B] = \sum \theta_i e_i$ where e_i is a basis for the cohomology with integer periods. Since $H^2(\Sigma; \mathbb{Z}) = \mathbb{Z}$ for Σ closed and orientable we can define a set of integers $d_i = \int f^*(e_i)$. Thus for example if $\Sigma = \mathbb{P}^1$ then “instanton sectors” will be labeled by d_i and weighted by

$$\prod_j e^{2\pi i d_j \theta_j} \quad (19.35)$$

Exercise *The Gaussian model on a torus*

Repeat the computation we did for the particle on the ring but for a free scalar field $\phi : T^2 \rightarrow T^n$ where the worldsheet and target space tori are provided with flat fields.

The analog of the Chern-Simons term is the “ B -field amplitude”

$$\exp[i \int B_{mn} d\phi^m \wedge d\phi^n] \quad (19.36)$$

where B_{mn} is a constant real antisymmetric matrix.

This is a long computation, but one of great importance in string theory.

19.4 MORE EXAMPLES

Example: The $O(3)$ nonlinear sigma model, also known as the CP^1 nonlinear sigma model in two dimensions. Here we have a field $\vec{n}$ defined on $\mathbb{R}^2$. The field satisfies a constraint $\vec{n}^2 = 1$ and hence we have a nonlinear sigma model with target S^2 . Finite action field configurations fall into components labelled by $\pi_2(S^2) = \mathbb{Z}$, and as we have seen these are measured by the degree (16.19).

This gives a topological term we can add to the action:

$$S = \frac{1}{\lambda^2} \int \|dn\|^2 + i\theta \deg(n) \quad (19.37)$$

In this model it is known that θ has important effects on the quantum physics:

1. At $\theta = 0$ the spectrum consists of an $O(3)$ triplet of massive particles.
 2. At $\theta = \pi$ the spectrum consists of an $SU(2)$ doublet of *massless* particles.
- $\theta = 0, \pi$ are distinguished because these are the two values which preserve CP .

References:

1. A classic review: Field Theory Methods And Quantum Critical Phenomena. Ian Affleck, Lectures given at Summer School on Fields, Strings and Critical Phenomena, Les Houches, 1988.
2. Another review: arXiv:0803.1593
3. Recently the issue of how one should sum over instanton sectors in the CP^1 model has been reexamined in an interesting paper of N. Seiberg. arXiv:1005.0002.

20. Sources

Notions from point-set topology: I will occasionally use various notions from point-set topology without definition. You can look these up in, for example:

1. J. Marsden, Elementary Classical Analysis
 2. J. Munkres, *Topology, a first course*, Prentice-Hall 1975
 3. J. Kelley, General Topology, Springer Verlag
 4. Nice online notes by A. Hatcher are at <http://www.math.cornell.edu/hatcher/Top/TopNotes.pdf>. (Hatcher also has other nice sets of notes on algebraic topology and K-theory.)
 5. M.A. Armstrong, *Basic Topology*, Springer Verlag
- and many many other places.

Basic algebraic topology can be found in

1. W.S. Massey, *Algebraic Topology: An Introduction*, Springer GTM 56
2. W.S. Massey,
3. R. Bott and L. Tu, An absolute classic. Highly recommended.
4. Bredon,
5. A. Hatcher, *Algebraic Topology*, Cambridge Univ. Press. 2002

More advanced:

1. Spanner,
2. Whitehead,

For physicists:

1. A.S. Schwarz, *Quantum Field Theory and Topology*,
2. M. Nakahara, *Geometry, Topology, and Physics*, Institute of Physics Publishing,

1995

Manifolds and differential topology:

1. M. Spivak, *Differential Geometry* vol. 1.
2. Bott, vol. 1, pp. 1 - 33
3. Bredon, ch. 1, sec. 13.
4. Dubrovin-Fomenko-Novikov, vol. III, sec. 1.4.
5. Guilleman and Pollack,
2. Nakahara, Sec. 2.2, 5.3-5.5
3. Isham, pp. 6-53
4. Schwarz, ch. 5
5. Helgason, ch. 1
6. Hawking and Ellis, ch. 1
7. Flanders, *Differential forms*
8. Eguchi, Gilkey, and Hanson, section 2
9. Boothby, Differentiable Manifolds
10. Thirring, Course in math physics.

Good references for topological sectors in physics, discussed only briefly above are:

1. S. Coleman, "Classical lumps and their quantum descendents," in *Aspects of Sym-*

metry

2. S. Coleman, "The magnetic monopole: 50 years later," Erice lectures.
3. L. Michel, Rev. Mod. Phys. **52** (1980) 617
4. D. Mermin, Rev. Mod. Phys. **51** (1979) 591
5. Nakahara, ch. 4 for many good examples.

Good treatments of solitons in field theory and string theory:

1. Callan, Harvey, Strominger, hep-th/9112030
2. David Tong, "TASI lectures on solitons: Instantons, monopoles, vortices and kinks,"

hep-th/0509216

3. Clifford V. Johnson, *D-branes* Cambridge, USA: Univ. Pr. (2003) 548 p.
4. L.D. Faddeev and L.A. Takhtadjan, *Hamiltonian Methods in the Theory of Solitons*

Springer

5. R. Rajaraman, *Solitons and Instantons, Volume 15: An Introduction to Solitons and Instantons in Quantum Field Theory*