1

Introduction

Human–like biomechanics is a modern scientific approach to *human–like motion dynamics and control*. Its *human perspective* has been developed in the work of the present authors (see [Iva91, ILI95, IS01, IP01b, IP01a, Iva02, Iva04, Iva05, IB05, PI03, PI04]). The dynamics of human motion is extremely complex, multi–dimensional, highly nonlinear and hierarchical. Human skeleton has more than two hundred rigid bones, connected by rotational joints, witch have up to three axes of rotation. Nevertheless, in classical biomechanics the main analytical tool was *translational vector geometry* (see Figure 1.1). The skeleton is driven by a synergistic action of its 640 skeletal muscles. Each of these muscles has its own *excitation* and *contraction* dynamics, in which neural action potentials are transformed into muscular force vectors (see [Hat77a, Hat77b, Hat78]).

On the other hand, *robotic approach* to human–like motion dynamics and control has been developed in the last tree decades in the work of M. Vukobratovic (and his collaborators). He started in the early 1970s with pioneering papers on synthesis, control and stability of biped gait [VJ69, VJF70, VFJ70], followed by mathematical models of locomotion robots and anthropomorphic mechanisms [VS72, VS73, Vuk75, Vuk78]. later, in 1980s, he developed the scientific fundamentals of robotics in seven volumes of the Springer–Verlag book series [VP82, VS82, VK85a, VK85b, VSK85, VP85, VBS89]. In recent years, within the realm of mature robotic science, Vukobratovic has been revising the study of anthropomorphism and intelligence of humanoid robots [VPM03, KV03a, KV03b, VB04, VPR, VPT04, VAB04, VBB05, PV05, RV05].

In this introductory Chapter we introduce the reader to the subject of modern human–like biomechanics, presenting its local (tensorial) language as well as its global (functorial) one. The objective here is to introduce our *covariant force law* (see subsection A.1.4 in Appendix),

$$F_i = mg_{ij}a^j, \qquad \text{that 'in plain English' reads :}$$

Force 1–form–field = Mass distribution $\times$ Acceleration vector–field

1.1 Local Tensorial Language of Human–Like Biomechanics

The local tensorial language of human–like biomechanics can be introduced in the following way. As it is pointed out in Appendix, the acceleration vector *is not* an ordinary time derivative of the velocity vector; 'even worse', the force, which is a paradigm of a vector in statics and engineering vector mechanics, *is not* a vector at all. The acceleration vector is the absolute time derivative of the velocity vector, while the force is a differential one–form. Proper description of these ideas is called *geometrodynamics*.[1]

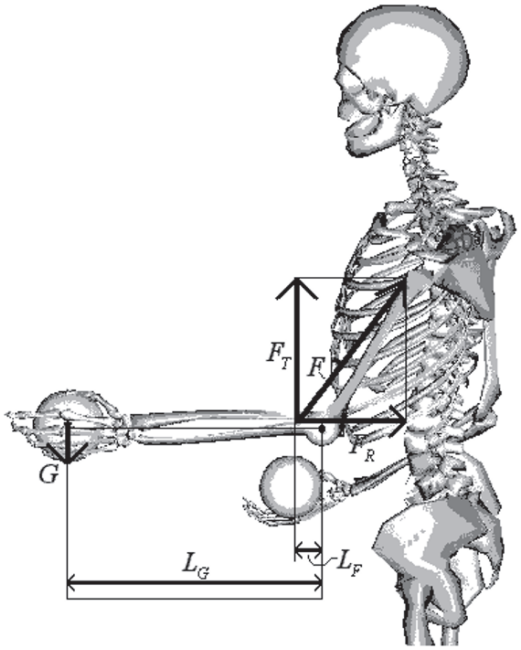


Fig. 1.1. Force vectors and their corresponding lever arms in classical translational biomechanics.

1.1.1 Classical Translational Biomechanics

Classical biomechanics consists of inverse and forward dynamics. *Inverse dynamics* is the commonly used technique used to gain insight into the net summation of all muscle activity at each joint. In this method, the joint forces

[1] Term *geometrodynamics* was coined by John A. Wheeler from Princeton.

and torques are calculated from a prescribed movement. Since the segmental movements, in contrast to the internal forces, can be measured, this method is commonly applied for the analysis of measured movements. A full kinematic description obtained from *motion capture* of marker positions is sufficient to get an inverse solution; however, motion capture is often combined with output from other sensors, including *force plates*, in order to improve the precision of the estimated joint loads [Dar]. On the other hand, the rarely used *forward dynamics* takes joint forces and torques as input to simulate translational and rotational motion. This paper focuses on the basic principles of rotational forward dynamics.

Most of the classical biomechanics is based on the *Newton's second law*, which states that a *conservative particle* of mass $m > 0$ immersed in a *potential* $V(q)$ moves along a *curve* $q^i(t)$ in *Euclidean 3D space* $\mathbb{R}^3$, in such a way that the *Newtonian equation of motion* is satisfied

$$m\ddot{q}^i = -\operatorname{grad} V(q), \tag{1.1}$$

for $q^i = \{q^1 \equiv x,\, q^2 \equiv y,\, q^3 \equiv z\} \in \mathbb{R}^3$.

Now, if we introduce the translational (or, the so–called 'linear') momenta $p_i = m\dot{q}^i$ and the total energy $H(q,p) = \frac{1}{2m}\|p\|^2 + V(q)$, (where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^3$), then the second Newton's law of motion is equivalent to *translational Hamiltonian equations of motion* (see [AMR88, Arn89, MR99]):

$$\dot{q}^i = \frac{\partial H}{\partial p_i} \equiv \partial_{p_i} H, \qquad \dot{p}_i = -\frac{\partial H}{\partial q^i} \equiv \partial_{q^i} H, \qquad (i = 1, \ldots, n). \tag{1.2}$$

One proceeds to study this system of first–order equations for a general $H(q,p)$. To do this, we introduce the matrix $J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$, where I is the 3×3 identity, and note that the equations become $\dot{\xi} = J \cdot \operatorname{grad} H(\xi)$, where $\xi = (q,p)$.

Set $X_H = J \cdot \operatorname{grad} H$. Then $\xi(t)$ satisfies Hamilton's equations iff $\xi(t)$ is an integral curve of X_H, that is, $\dot{\xi}(t) = X_H(\xi(t))$ [AMR88].

The space $\mathbb{R}^3 \times \mathbb{R}^3$ of the ξ's is called the *phase space*. For a system of n particles we would use $\mathbb{R}^{3n} \times \mathbb{R}^{3n}$.

The conservative Hamiltonian system (1.2) is characterized by conservation of energy, and consequently momentum; the classical example of the later is reactive recoil of a gun after shooting.

The left, $\dot{q}-$equation in (1.2) is called the 'Velocity equation', while the right, $\dot{p}-$equation is called the 'Force equation'. Combined, they give the biomechanical sense of the celebrated Hill's 'Force–Velocity relation' ([Hil38]). All biologically essential non-conservative forces, like neuro–muscular servo–drives, tendon elasticities and joint dampings, are to be added to the force equation. This is probably the most plausible way of presenting the *translational biomechanics*.

1.1.2 Calculus of Geometric Objects

Following the trend that has prevailed in physical sciences for several decades (see, e.g., [MTW73]), we state here the leading idea of this book: *every biomechanical quantity can be described by a geometric object; all biomechanical laws can be expressed as geometric relationships between these geometric objects.* This trend has its mathematical beginnings in the Erlanger Program of Felix Klein and entered the physical sciences in the form of the 'principle of general covariance' of Albert Einstein: *physical laws must be independent of any particular coordinate systems if they are to be valid.* A study of the consequences of this requirement leaded, at first, to classical tensor calculus, and subsequently, to its modern development, *calculus of coordinate–free geometric objects*, the most natural machinery for analysis of biomechanical systems. We assume that our reader is familiar with classical tensor calculus, while the modern calculus of coordinate–free geometric objects will be developed as we proceed in our study of human–like biomechanics.

In biomechanics, the most important geometric objects are Riemannian metric and curvature tensors (see Appendix), as well as several exterior differential forms.

Metric Tensor – the Core of Geometrodynamics

The act of measurement in curved spaces is performed by the use of Riemannian metrics. The general result is this: at each point of any smooth *Riemannian manifold* there exists a geometric object called the *metric tensor* g, usually given by its *covariant components* g_{ij}. However, it is easier to start with our familiar ordinary Euclidean three–dimensional space with rectilinear Cartesian axes defined on it.

In any case, the metric tensor $g = (g_{ij})$ defines a linear symmetrical machine $g(\cdot, \cdot)$ with two input slots for the insertion of two vectors, producing a real number as an output. It can be used for calculating the *scalar product*

$$v \cdot w = g(v, w) = g_{ij} v^i w^j$$

of two different vectors $v = (v^i)$ and $w = (w^i)$, or the square length

$$g(v, v) = v^2$$

of a single vector v.

In particular, in *Euclidean 3D space* $\mathbb{R}^3$, with the coordinate basis

$$\{e_1 = dx,\ e_2 = dy,\ e_3 = dz\}$$

along the standard *Cartesian axes* $\{x, y, z\}$, the metric tensor $g = (g_{ij})$ is defined by

$$g = g(e_i, e_j) = e_i \cdot e_j,$$

while the vector $v = (v^i)$ is expressed in components as

$$v = v^1 e_1 + v^2 e_2 + v^3 e_3 = v^i e_i$$

(Einstein's summation convention is always in place).

At this point we need to emphasize that more fundamental than the components of a tensor (or a vector) is the tensor (respectively vector) itself, a geometric object with a physical or biomechanical meaning independent of all coordinates. For example, velocity of a particle moving in $\mathbb{R}^3$ is a *tangent vector* to the trajectory of the particle, defined locally as a derivative of the trajectory at a certain point. However, coordinates necessarily enter the scene when numerical analysis is required.

Classical Tensor Dynamics in Brief

Now we switch from geometry to dynamics, intending to show that they are actually the same thing, sometimes called *geometrodynamics*. Recall that a material system is regarded from the dynamical standpoint as a collection of particles which are subject to interconnections and constraints of various kinds (e.g., a rigid body is regarded as a number of particles rigidly connected together so as to remain at invariable distances from each other). The *number of independent coordinates which determine the configuration of a dynamical system completely* is called the *number of degrees of freedom* (DOF) of the system. In other words, this number, n, is the *dimension of the system's configuration manifold*. This viewpoint is the core of our geometric biomechanics.

For simplicity, let us suppose that we have a dynamical system with three DOF (e.g., a particle of mass M, or a rigid body of mass M with one point fixed); generalization to n DOF, with N included masses M_α, is straightforward. The configuration of our system at any time is then given by three coordinates $\{q^1, q^2, q^3\}$. As the coordinates change in value the dynamical system changes its configuration. Obviously, there is an infinite number of sets of independent coordinates which will determine the configuration of a dynamical system, but since the position of the system is completely given by any one set, these sets of coordinates must be functionally related. Hence, if $\bar{q}^i$ is any other set of coordinates, these quantities must be connected with q^i by formulae of the type

$$\bar{q}^i = \bar{q}^i(q^i), \qquad (i = 1, ..., n(= 3)). \tag{1.3}$$

Relations (1.3) are the equations of transformation from one set of dynamical coordinates to another and, in a standard tensorial way (see Appendix, as well as e.g., [McC60, SS78, BL81, LR89, LC03], although we recommend [MTW73]), we can define tensors relative to this coordinate transformation. The generalized coordinates q^i, $(i = 1, ..., n)$ constitute the system's *configuration manifold*.

In particular, in our ordinary Euclidean 3−dimensional (3D) space $\mathbb{R}^3$, the ordinary Cartesian axes are $x^i = \{x, y, z\}$, and the *system's center of mass* (COM) is given by

$$C^i = \frac{M_\alpha x^i_\alpha}{\sum_{\alpha=1}^{N} M_\alpha},$$

where Greek subscript α labels the masses included in the system. If we have a continuous distribution of matter $V = V(M)$ rather than the discrete system of masses M_α, all the α−sums should be replaced by volume integrals, the element of mass dM taking the place of M_α,

$$\sum_{\alpha=1}^{N} M_\alpha \Rightarrow \iiint_V dM.$$

An important quantity related to the system's COM is the double symmetric contravariant tensor

$$I^{ij} = M_\alpha x^i_\alpha x^j_\alpha, \tag{1.4}$$

called the *inertia tensor*, calculated relative to the origin O of the Cartesian axes $x^i_\alpha = \{x_\alpha, y_\alpha, z_\alpha\}$. If we are given a straight line through O, defined by its unit vector λ^i, and perpendiculars p_α are drawn from the different particles on the line λ^i, the quantity

$$I(\lambda^i) = M_\alpha p^2_\alpha$$

is called the *moment of inertia* around λ^i. The moment of inertia $I(\lambda^i)$ can be expressed through inertia tensor (1.4) as

$$I(\lambda^i) = (I g_{ij} - I_{ij})\lambda^i \lambda^j,$$

where g_{ij} is the system's *Euclidean 3D metric tensor* (as defined above), $I = g_{ij} I^{ij}$, and $I_{ij} = g_{rm} g_{sn} I^{mn}$ is the *covariant inertia tensor*. If we now consider the quadric Q whose equation is

$$(I g_{ij} - I_{ij}) x^i x^j = 1, \tag{1.5}$$

we find that the moment of inertia around λ^i is $1/R$, where R is the radius vector of Q in the direction of λ^i. The quadric Q defined by relation (1.5) is called the *ellipsoid of inertia* at the origin O. It has always three principal axes, which are called the *principal axes of inertia* at O, and the planes containing them in pairs are called the *principal planes of inertia* at O. The principal axes of inertia are given by the equations

$$(I g_{ij} - I_{ij})\lambda^j = \theta \lambda^i,$$

where θ is a root of the determinant equation

$$|(I - \theta) g_{ij} - I_{ij}| = 0.$$

More generally, if we suppose that the points of our dynamical system are referred to rectilinear Cartesian axes x^i in a Euclidean $n-$dimensional (nD) space $\mathbb{R}^n$, then when we are given the time and a set of *generalized coordinates* q^i we are also given all the points x^i of the dynamical system, as the system is determined uniquely. Consequently, the x^i are functions of q^i and possibly also of the time, that is,

$$x^i = x^i(q^i, t).$$

If we restrict ourselves to the *autonomous dynamical systems* in which these equations do not involve t, i.e.,

$$x^i = x^i(q^i), \tag{1.6}$$

then differentiating (1.6) with respect to the time t gives

$$\dot{x}^i = \frac{\partial x^i}{\partial q^j}\dot{q}^j. \tag{1.7}$$

The quantities $\dot{q}^i$, which form a vector with reference to coordinate transformations (1.3), we shall call the *generalized velocity vector*. We see from (1.7) that when the generalized velocity vector is given we know the velocity of each point of our system. Further, this gives us the system's *kinetic energy*,

$$E_{kin} = \frac{1}{2}M_\alpha g_{mn}\dot{x}^m_\alpha \dot{x}^n_\alpha = \frac{1}{2}M_\alpha g_{mn}\frac{\partial x^m_\alpha}{\partial q^i}\frac{\partial x^n_\alpha}{\partial q^j}\dot{q}^i\dot{q}^j. \tag{1.8}$$

Now, if we use the Euclidean metric tensor g_{ij} to define the *material metric tensor* G_{ij}, including the distribution of all the masses M_α of our system, as

$$G_{ij} = M_\alpha g_{mn}\frac{\partial x^m_\alpha}{\partial q^i}\frac{\partial x^n_\alpha}{\partial q^j}, \tag{1.9}$$

the kinetic energy (1.8) becomes a *homogenous quadratic form* in the generalized system's velocities $\dot{q}^i$,

$$E_{kin} = \frac{1}{2}G_{ij}\dot{q}^i\dot{q}^j. \tag{1.10}$$

From the transformation relation (1.9) we see that the material metric tensor G_{ij} is symmetric in i and j. Also, since E_{kin} is an invariant for all transformations of generalized coordinates, from (1.10) we conclude that G_{ij} is a double symmetric tensor. Clearly, this is the central quantity in classical tensor system dynamics. We will see later, that G_{ij} defines the Riemannian geometry of the system dynamics. For simplicity reasons, G_{ij} is often denoted by purely geometric symbol g_{ij}, either assuming or neglecting the material properties of the system.

Now, let us find the equations of motion of our system. According to the *D'Alembert's Principle of virtual displacement*s, the equations of motion in Cartesian coordinates x^i in $\mathbb{R}^n$ are embodied in the single tensor equation

$$g_{mn}(M_\alpha \ddot{x}^m_\alpha - X^m_\alpha)\delta x^n_\alpha = 0, \tag{1.11}$$

where X^i_α is the total force vector acting on the particle M_α, while δx^i_α is the associated *virtual displacement* vector, so that the product $g_{ij}X^i_\alpha \delta x^j_\alpha$ is the *virtual work* of the system, and we can neglect in X^i_α all the internal or external forces which do not work in the displacement δx^i_α. If we give the system a small displacement compatible to with the constraints of the system, we see that this displacement may be effected by giving increments δq^i to the generalized coordinates q^i of the system, and these are related to the δx^i in accordance with the transformation formulae $\delta x^i_\alpha = \frac{\partial x^i_\alpha}{\partial q^j}\delta q^j$.

Furthermore, in this displacement the internal forces due to the constraints of the system will do no work, since these constraints are preserved, and consequently only the external forces will appear in (1.11), so it becomes

$$g_{mn}\left[M_\alpha \frac{d}{dt}\left(\frac{\partial x^m_\alpha}{\partial q^j}\dot{q}^j\right)\frac{\partial x^n_\alpha}{\partial q^i} - X^m_\alpha \frac{\partial x^n_\alpha}{\partial q^i}\right]\delta q^i = 0. \tag{1.12}$$

Now, using (1.8–1.10), we derive

$$M_\alpha g_{mn}\frac{d}{dt}\left(\frac{\partial x^m_\alpha}{\partial q^j}\dot{q}^j\right)\frac{\partial x^n_\alpha}{\partial q^i} = \frac{d}{dt}(G_{ij}\dot{q}^j) - \frac{1}{2}\frac{\partial G_{st}}{\partial q^i}\dot{q}^j\dot{q}^k = \frac{d}{dt}\left(\frac{\partial E_{kin}}{\partial \dot{q}^i}\right) - \frac{\partial E_{kin}}{\partial q^i}.$$

Also, if we put

$$F_i = g_{mn}X^m_\alpha \frac{\partial x^n_\alpha}{\partial q^i},$$

we get

$$F_i\delta q^i = g_{mn}X^m_\alpha \delta x^n_\alpha = \delta W, \tag{1.13}$$

where δW is the virtual work done by the external forces in the small displacement δq^i, which shows that F_i is the covariant vector, called the *generalized force vector*. Now (1.12) takes the form

$$\left[\frac{d}{dt}\left(\frac{\partial E_{kin}}{\partial \dot{q}^i}\right) - \frac{\partial E_{kin}}{\partial q^i} - F_i\right]\delta q^i = 0.$$

Since the coordinates q^i are independent this equation is true for all variations δq^i and we get as a final result the *covariant Lagrangian equations of motion*,

$$\frac{d}{dt}\left(\frac{\partial E_{kin}}{\partial \dot{q}^i}\right) - \frac{\partial E_{kin}}{\partial q^i} = F_i.$$

If the force system is conservative and E_{pot} is the system's *potential energy* given by

$$F_i = -\frac{\partial E_{pot}}{\partial q^i},$$

then, using (1.13) the Lagrangian equations take the standard form

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^i}\right) = \frac{\partial L}{\partial q^i}, \tag{1.14}$$

where the *Lagrangian function* $L = L(q, \dot{q})$ of the system is given by $L = E_{kin} - E_{pot}$ (since E_{pot} does not contain $\dot{q}^i$).

Now, the kinetic energy E_{kin} of the system, given by quadratic form (1.10), is always positive except when $\dot{q}^i$ is zero in which case E_{kin} vanishes. In other words, the quadratic form (1.10) is positive definite. Consequently, we can always find the *line* (or *arc*) *element*, defined by

$$ds^2 = G_{ij} dq^i dq^j. \tag{1.15}$$

A manifold in which ds^2 is given by relation of the type of (1.15), geometrically with g_{ij} instead of G_{ij}, is called a *Riemannian manifold*.

Riemannian Curvature Tensor

Every Riemannian manifold is characterized by the *Riemann curvature tensor*. In physical literature (see, e.g., [MTW73]) it is usually introduced through the *Jacobi equation of geodesic deviation*, showing the acceleration of the relative separation of nearby geodesics (the shortest, straight lines on the manifold). For simplicity, consider a sphere of radius a in $\mathbb{R}^3$. Here, Jacobi equation is pretty simple,

$$\frac{d^2\xi}{ds^2} + R\xi = 0,$$

where ξ is the *geodesic separation vector* (the so–called Jacobi vector–field), s denotes the geodesic arc parameter given by (1.15) and $R = 1/a^2$ is the *Gaussian curvature* of the surface.

In case of a higher–dimensional manifold M, the situation is naturally more complex, but the main structure of the Jacobi equation remains similar,

$$\frac{D^2\xi}{ds^2} + R(u, \xi, u) = 0,$$

where D denotes the covariant derivative and $R(u, \xi, u)$ is the curvature tensor, a three–slot linear machine. In components defined in a local coordinate chart (x^i) on M, this equation reads

$$\frac{D^2\xi^i}{ds^2} + R^i_{jkl}\frac{dx^j}{ds}\xi^k\frac{dx^l}{ds} = 0,$$

where R^i_{jkl} are the components of the Riemannian curvature tensor.

Exterior Differential Forms

Recall that *exterior differential forms* are a special kind of antisymmetrical covariant tensors (see, e.g., [DRh84, Fla63]). Such tensor–fields arise in many

applications in physics, engineering, and differential geometry. The reason for this is the fact that the classical vector operations of **grad**, **div**, and **curl** as well as the theorems of Green, Gauss, and Stokes can all be expressed concisely in terms of differential forms and the main operator acting on them, the exterior derivative d. Differential forms inherit all geometric properties of the general tensor calculus and add to it their own powerful geometric, algebraic and topological machinery (see Figures 1.2 and 1.3). Differential $p-$forms formally occur as *integrands* under ordinary integral signs in $\mathbb{R}^3$:

- a *line integral* $\int P\,dx + Q\,dy + R\,dz$ has as its integrand the *one–form* $\omega = P\,dx + Q\,dy + R\,dz$;
- a *surface integral* $\iint A\,dydz + B\,dzdx + C\,dxdy$ has as its integrand the *two–form* $\alpha = A\,dydz + B\,dzdx + C\,dxdy$;
- a *volume integral* $\iiint K\,dxdydz$ has as its integrand the *three–form* $\lambda = K\,dxdydz$.

By means of an *exterior derivative* d, a *derivation* that transforms $p-$forms into $(p+1)-$forms, these geometric objects generalize ordinary vector differential operators in $\mathbb{R}^3$:

- a *scalar function* $f = f(x)$ is a zero–form;
- its *gradient* df, is a one–form[2]

$$df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial z}dz;$$

- a *curl* $d\omega$, of a one–form ω above, is a two–form

$$d\omega = \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right)dydz + \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right)dzdx + \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)dxdy;$$

- a *divergence* $d\alpha$, of the two–form α above, is a three–form

$$d\alpha = \left(\frac{\partial A}{\partial x} + \frac{\partial B}{\partial y} + \frac{\partial C}{\partial z}\right)dxdydz.$$

Now, although visually intuitive, our Euclidean 3D space $\mathbb{R}^3$ is not sufficient for thorough biomechanical analysis. The fundamental concept of a *smooth manifold*, locally topologically equivalent to the Euclidean nD space $\mathbb{R}^n$, is required (with or without Riemannian metric tensor defined on it). In general, a proper definition of exterior derivative d for a $p-$form β on a smooth manifold M, includes the *Poincaré lemma*: $d(d\beta) = 0$, and validates the general integral *Stokes formula*

[2] We use the same symbol, d, to denote both ordinary and exterior derivation, in order to avoid extensive use of the boldface symbols. It is clear from the context which derivative (differential) is in place: exterior derivative operates only on differential forms, while the ordinary differential operates mostly on coordinates.

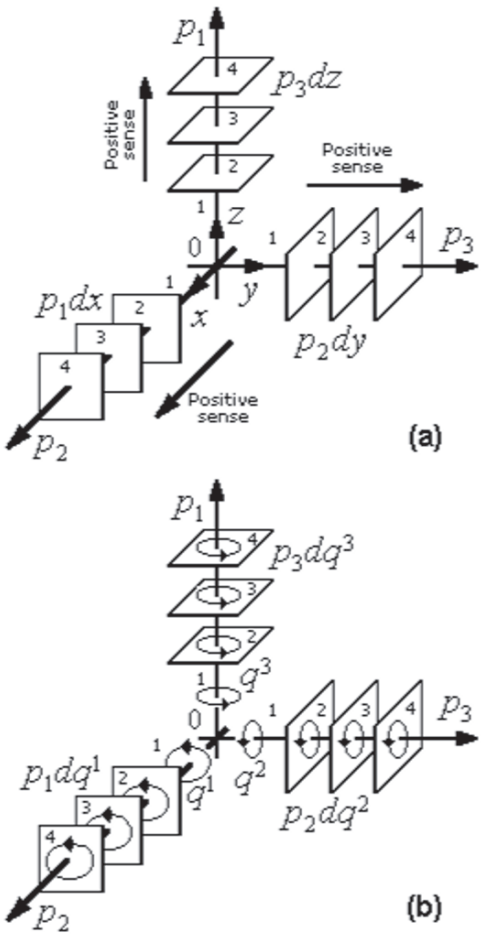


Fig. 1.2. Basis vectors and one-forms in Euclidean $\mathbb{R}^3-$space: (a) Translational case; and (b) Rotational case.

$$\int_{\partial M} \beta = \int_M d\beta,$$

where M is a $p-$dimensional *manifold with a boundary* and ∂M is its $(p-1)-$dimensional *boundary*, while the integrals have appropriate dimensions.

A $p-$form β is called *closed* if its exterior derivative is equal to zero,

$$d\beta = 0.$$

From this condition one can see that the closed form (the *kernel* of the exterior derivative operator d) is conserved quantity. Therefore, closed $p-$forms possess certain invariant properties, physically corresponding to the conservation laws.

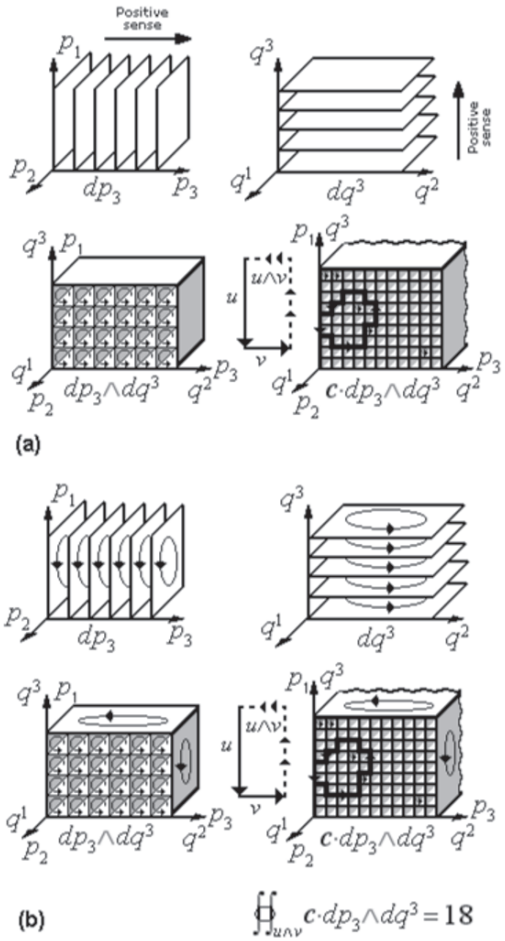


Fig. 1.3. Fundamental two–form and its flux in $\mathbb{R}^3$: (a) Translational case; (b) Rotational case. In both cases the flux through the plane $u\wedge v$ is defined as $\int\int_{u\wedge v} c\, dp_i dq^i$ and measured by the number of tubes crossed by the circulation oriented by $u \wedge v$.

A $p-$form β that is an exterior derivative of some $(p-1)-$form α,

$$\beta = d\alpha,$$

is called *exact* (the *image* of the exterior derivative operator d). By Poincaré Lemma, exact forms prove to be closed automatically,

$$d\beta = d(d\alpha) = 0.$$

Similarly to the components of a 3D vector v defined above, a one–form θ defined on an nD manifold M can also be expressed in components, using

the coordinate basis $\{dx^i\}$ along the local nD coordinate chart $\{x^i\} \in M$, as

$$\theta \;=\; \theta_i \, dx^i.$$

Now, the components of the exterior derivative of θ are equal to the components of its *commutator* defined on M by

$$d\theta = \omega_{ij} \, dx^i \, dx^j,$$

where the components of the *form commutator* ω_{ij} are given by

$$\omega_{ij} \;=\; \left(\frac{\partial \theta_i}{\partial x^i} \;-\; \frac{\partial \theta_i}{\partial x^j} \right).$$

The space of all smooth $p-$forms on a smooth manifold M is denoted by $\Omega^p(M)$. The *wedge*, or *exterior product* of two differential forms, a $p-$form $\alpha \in \Omega^p(M)$ and a $q-$form $\beta \in \Omega^q(M)$ is a $(p+q)-$form $\alpha \wedge \beta$. For example, if $\theta = a_i dx^i$, and $\eta = b_j dx^j$, their wedge product $\theta \wedge \eta$ is given by

$$\theta \wedge \eta = a_i b_j dx^i dx^j,$$

so that the coefficients $a_i b_j$ of $\theta \wedge \eta$ are again smooth functions, being polynomials in the coefficients a_i of θ and b_j of η. The exterior product $\wedge$ is related to the exterior derivative $d : \Omega^p(M) \to \Omega^{p+1}(M)$, by

$$d(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^p \alpha \wedge d\beta.$$

Another important linear operator is the *Hodge star* $* \;:\; \Omega^p(M) \;\to\; \Omega^{n-p}(M)$, where n is the dimension of the manifold M. This operator depends on the inner product (i.e., Riemannian metric) on M and also depends on the orientation (reversing orientation will change the sign). For any $p-$forms α and β,

$$** \alpha = (-1)^{p(n-p)} \alpha, \quad \text{and} \quad \alpha \wedge *\beta = \beta \wedge *\alpha.$$

Hodge star is generally used to define *dual* $(n-p)-$*forms* on nD smooth manifolds.

For example, in $\mathbb{R}^3$ with the ordinary Euclidean metric, if f and g are functions then (compare with the 3D forms of gradient, curl and divergence defined above)

$$\begin{aligned}
df &= \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial z} dz, \\
*df &= \frac{\partial f}{\partial x} dydz + \frac{\partial f}{\partial y} dzdx + \frac{\partial f}{\partial z} dxdy, \\
df \wedge *dg &= \left(\frac{\partial f}{\partial x} \frac{\partial g}{\partial x} + \frac{\partial f}{\partial y} \frac{\partial g}{\partial y} + \frac{\partial f}{\partial z} \frac{\partial g}{\partial z} \right) dxdydz = \Delta f \, dxdydz,
\end{aligned}$$

where Δf is the *Laplacian* on $\mathbb{R}^3$. Therefore the three–form $df \wedge *dg$ is the Laplacian multiplied by the volume element, which is valid, more generally, in any local orthogonal coordinate system in any smooth domain $U \in \mathbb{R}^3$.

The subspace of all closed $p-$forms on M we will denote by $Z^p(M) \subset \Omega^p(M)$, and the sub-subspace of all exact $p-$forms on M we will denote by $B^p(M) \subset Z^p(M)$. Now, the quotient space

$$H^p(M) = \frac{Z^p(M)}{B^p M} = \frac{\mathrm{Ker}\left(d : \Omega^p(M) \to \Omega^{p+1}(M)\right)}{\mathrm{Im}\,(d : \Omega^{p-1}(M) \to \Omega^p(M))}$$

is called the pth *De Rham cohomology group* (or vector space) of a manifold M. Two $p-$forms α and β on M are equivalent, or belong to the same *cohomology class* $[\alpha] \in H^p(M)$, if their difference equals $\alpha - \beta = d\theta$, where θ is a $(p-1)-$form on M.

1.1.3 Lagrangian Action and Feynman Path Integral

Recall that all the fundamental laws of classical physics can be understood in terms of one mathematical construct, the *action principle*, as well as all the fundamental laws of quantum physics can be understood in terms of associated construct, the *Feynman path integral* (see, e.g., [Ram90]).

In 1746 Maupertuis formulated the Principle of Least Action, which is all too commonly credited to one of the three great mathematicians, Euler, Lagrange, and Hamilton, who further developed it. This principle is one of the greatest generalizations in all physical science, although not fully appreciated until the advent of quantum mechanics and Feynman path integral in the mid 20th century.

In particular, all Newtonian particle mechanics is contained in the *Hamilton's principle* of *least action*, which demands that the true trajectory $x = x(t)$ of a particle is that function which *minimizes the action* $S[x(t)]$, given as a temporal integral of the autonomous *Lagrangian function* $L = L(x, \dot{x})$,

$$S[x] = \int_{t_0}^{t_1} L(x, \dot{x})dt \longrightarrow \min, \qquad \text{or,} \qquad \delta S[x] = 0,$$

where the second expression reads: 'variation of the action equals zero' and implies using techniques from the *calculus of variations* (see e.g., [For60]).

Now, associated to the least action principle is the path integral, or Feynman's *sum–over–histories*.[3] While Nature's command for the classical particle is: "Follow the path of least action," to the elementary particle it commands: "Explore all possible paths!"

[3] Here we quote F. Dyson: "Dick Feynman told me about his sum–over–histories version of quantum mechanics. "The electron does anything it likes," he said. "It just goes in any direction at any speed, forward or backward in time, however it likes, and then you add up the amplitudes and it gives you the wave–function." I said to him, "You're crazy." But he wasn't."

According to this general *action principle – path integral* formalism (see section 3.3 as well as Appendix, section A.3), we first formulate the acceptable *action functional* (AF, denoted by square brackets [...]), from which we derive *Euler–Lagrangian equations* of motion, and subsequently we perform the *quantization* of the system and find its *transition amplitude*, by evaluating the *associate path integral*.

In mechanics, the *Lagrangian action formalism* is a four–step algorithm with a purpose of finding the *true*, or *straight path* $x^i = x^i(t)$, $(i = 1, ..., N)$ of an N–degree–of–freedom autonomous dynamical system:

1. Formulate the *Lagrangian function* $L = L(x, \dot{x})$ of the system, e.g., as a function $L(x, \dot{x}) = E_{kin}(\dot{x}) - E_{pot}(x)$;
2. Write down the AF, as a temporal integral,
$$S[x] = \int_{t_0}^{t_1} L dt;$$
3. Formulate the *action principle*, as a vanishing variation of the AF,
$$\delta S[x] = \delta \int_{t_0}^{t_1} L dt = \int_{t_0}^{t_1} \delta L dt = 0,$$
with zero initial and final path variations, $\delta x(t_0) = \delta x(t_1) = 0$;
4. Derive the *Euler–Lagrangian equations* of motion,
$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}^i}\right) = \frac{\partial L}{\partial x^i}, \qquad (i = 1, ..., N),$$
using the vanishing *functional derivative*, $\frac{\delta S}{\delta x^i} = 0$, given by (using standard variational techniques)
$$\frac{\delta S}{\delta x^i} \equiv \frac{\partial L}{\partial x^i} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}^i}\right).$$

For illustration, take the simple example of a single *point particle* moving in Euclidean 3D space, with position vector $x^i = x^i(t)$ $(i = 1, 2, 3)$, at time t, within a time independent *potential field* $V(x^i)$. The corresponding AF is given by
$$S\left(\left[x^i\right], t_0, t_1\right) = \int_{t_0}^{t_1} dt \left(\frac{1}{2} m \frac{dx^i}{dt}\frac{dx^i}{dt} - V(x^i)\right). \tag{1.16}$$
The AF (1.16) is a *function* of the initial and final times, t_0 and t_1 (which we write $S(t_0, t_1)$), and at the same time a *functional of the path* $x^i(t)$ (which we write $S\left[x^i\right]$) for $t_0 < t_1$.[4] To build the AF, we consider a small deformation of the path,
$$x^i(t) \to x^i(t) + \delta x^i(t).$$

[4] For example, the length of a path is a functional of the path.

The $S-$response to the small path deformation is given by

$$S\left[x^i+\delta x^i\right]=\int_{t_0}^{t_1}dt\left(\frac{1}{2}m\frac{d(x^i+\delta x^i)}{dt}\frac{d(x^i+\delta x^i)}{dt}-V(x^i+\delta x^i)\right) \quad (1.17)$$
$$=S\left[x^i\right]+\int_{t_0}^{t_1}dt\,\delta x^i\left(-\partial_i V(x^i)-m\ddot{x}^i\right)+m\int_{t_0}^{t_1}dt\,\frac{d}{dt}\left(\delta x^i\dot{x}^i\right),$$

where $\partial_i\equiv\partial_{x^i}\equiv\frac{\partial}{\partial x^i}$, and

$$V(x^i+\delta x^i)=V(x^i)+\delta x^i\partial_i V(x^i).$$

According to the standard variation techniques, the last term in (1.17) is just a 'surface' term, which is usually eliminated by restricting the variations to paths which vanish at the end points,

$$\delta x^i(t_0)=\delta x^i(t_1)=0.$$

In this way, (1.17) becomes

$$S\left[x^i+\delta x^i\right]=S\left[x^i\right]+\int_{t_0}^{t_1}dt\,\delta x^i\frac{\delta S}{\delta x^i},$$

where the functional derivative $\frac{\delta S}{\delta x^i}$ is defined by

$$\frac{\delta S}{\delta x^i}=-(m\ddot{x}^i+\partial_i V(x^i)). \quad (1.18)$$

Therefore, the minimization of the AF (1.16) is, according to the Hamilton action principle, equivalent to the vanishing functional derivative, which leads to the equations of motion

$$\frac{\delta S}{\delta x^i}\equiv-(m\ddot{x}^i+\partial_i V(x^i))=0,\qquad \text{or,}\qquad m\ddot{x}^i=\partial_i V(x^i).$$

Note, however, that minimization of S only leads to a *class of possible paths*. Which of those is followed depends on the boundary conditions, given as initial values of positions $x^i(t_0)$ and velocities $\dot{x}^i(t_0)\equiv\frac{dx^i(t_0)}{dt}$.

More generally, in *field theory*, the four–step Lagrangian action formalism is the following algorithm:

1. Formulate the *Lagrangian density* $\mathcal{L}=\mathcal{L}(\varphi^i,\partial_\mu\varphi^i)$ of the system as a function of m field variables $\varphi^i=\varphi^i(x^\mu)$ and their first partial derivatives $\partial_\mu\varphi^i$ over the n system coordinates x^μ (e.g., 4 space–time coordinates);
2. Write down the AF, as an nD integral,

$$S[x]=\int\mathcal{L}(\varphi^i,\partial_\mu\varphi^i)\,dx,\qquad dx=\prod_{\mu=1}^{n}dx^\mu;$$

3. Formulate the *action principle*, as a vanishing variation of the AF,

$$\delta S[x] = \delta \int \mathcal{L} dx = \int \delta \mathcal{L} dx = 0,$$

with zero boundary field variations;
4. Derive the *Euler–Lagrangian equations* of motion, using the vanishing *functional derivative*, $\frac{\delta S}{\delta x^i} = 0$, given by

$$\frac{\delta S}{\delta \varphi^i} \equiv \frac{\partial \mathcal{L}}{\partial \varphi^i} - \partial_\mu \left(\frac{\partial \mathcal{L}}{\partial_\mu \varphi^i} \right).$$

Now, once we have an acceptable AF, we can formulate the associated Feynman path integral, according to the procedures developed in Appendix. In case of a single point particle, the path integral corresponding to the AF (1.16) is formally written as (we use the normal units with $\hbar = 1$; also, $\mathrm{i} \equiv \sqrt{-1}$)

$$\int \mathcal{D}[x]\, \mathrm{e}^{\mathrm{i}S([x],t_0,t_1)} = \int \mathcal{D}[x] \exp\left[\mathrm{i} \int_{t_0}^{t_1} dt \left(\frac{1}{2} m\, \dot{x}^i \dot{x}^i - V(x^i) \right) \right], \qquad (1.19)$$

where

$$\int \mathcal{D}[x] \approx \int \prod_{k=1}^{N} \frac{dx_k^i}{\sqrt{2\pi k dt}}$$

is the *Lebesgue integration over all possible complex–valued trajectories* $x^i = x^i(t)$ between t_0 and t_1, performed by splitting the time interval $[t_0, t_1]$ into N subintervals (see Appendix for details). Integral (1.19) represents the *transition amplitude* $\langle X_1^i | X_0^i \rangle$ for the particle 'jumping' from point $x^i(t_0)$ to point $x^i(t_1)$, where $X^i = X^i(t)$ is the *Hermitian position operator* corresponding to the coordinate x^i such that the boundary condition for x^i at point t_0 is X_0^i and x^i at point t_1 is X_1^i. The transition amplitude is then given by

$$\langle X_1^i | X_0^i \rangle = \int \mathcal{D}[x]\, \mathrm{e}^{\mathrm{i}S([x],t_0,t_1)}.$$

In case of external driving forces, $F_i = F_i(t)$, the AF (1.16) is expanded into

$$S([x], t_0, t_1) = \int_{t_0}^{t_1} dt \left(\frac{1}{2} m\, \dot{x}^i \dot{x}^i - V(x^i) + F_i(t)\, x^i(t) \right),$$

and the corresponding *forced transition amplitude* becomes (see, e.g., [Ram90])

$$\langle X_1^i | X_0^i \rangle_F = \int \mathcal{D}[x] \exp\left[\mathrm{i} \int_{t_0}^{t_1} dt \left(\frac{1}{2} m\, \dot{x}^i \dot{x}^i - V(x^i) + F_i(t)\, x^i(t) \right) \right].$$

For the derivation of the path integral, see Appendix. In Chapter 5 we will formulate the path–integral model for the neural control of human motion. In Chapter 6 we will use more general actions and path integrals to explore biophysics of electro–muscular stimulation.

1.1.4 Noether Theorem

Recall that the 'surface term' G of the general action functional (AF)

$$S\left(\left[x^i\right], t_0, t_1\right) = \int_{t_1}^{t_2} L(x^i, \dot{x}^i)\, dt, \qquad (i = 1, ..., N) \tag{1.20}$$

offers a connection between the *conservation laws* and the *invariants of the dynamical system*, governed by the celebrated *Noether theorem* (see [Ram90, AM78, Arn89, MR99]).

Noether's theorem relates pairs of basic ideas of physics, one being the invariance of the form that a physical law takes with respect to any (generalized) transformation that preserves the coordinate system (both spatial and temporal aspects taken into consideration), and the other being a conservation law of a physical quantity. Informally, Noether's theorem can be stated as: *There is a one-to-one correspondence between continuous symmetries of the laws of physics, and conservation laws in physics.* More precisely, yet still informal: *To every differentiable symmetry which is generated by local actions, there corresponds a conservation law, defining a conserved current, and vice versa.* The formal statement of the theorem derives an expression for the physical quantity that is conserved (and hence also defines it (actually, its current)), from the condition of invariance alone. For example: (i) Invariance of physical systems with respect to *translation* gives the law of *conservation of linear momentum* (when simply stated, it is just that the laws of physics don't vary with location in space); (ii) Invariance with respect to *rotation* gives law of *conservation of angular momentum*; (iii) Invariance with respect to *time* gives the well known law of *conservation of energy*, etc.

To get some 'feeling' for the Noether's theorem, recall that the *Lagrangian equations* corresponding to the action (1.20) read

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}^i}\right) = \frac{\partial L}{\partial x^i}.$$

These equations signify that if the expression on the r.h.s *is zero*, $\frac{\partial L}{\partial x^i} = 0$, meaning that L is symmetrical over the coordinates $x^i = x^i(t)$, then the *rate of change* of the expression in parentheses on the l.h.s *is also zero*, $\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}^i}\right) = 0$, and therefore, the N *generalized momenta*, $p_i = \frac{\partial L}{\partial \dot{x}^i}$, are *conserved quantities*.

Despite the fact that the classical Lagrangian equation, $\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}^i}\right) = \frac{\partial L}{\partial x^i}$, is essentially an explicit statement of this proposition, it seems not to have been discussed and formalized as a theorem until 1918, by Emmy Noether (1882–1935), so it is now called Noether's Theorem. This theorem was praised by Einstein as a piece of "penetrating mathematical thinking". It is now a standard workhorse in theoretical physics.

More precisely, let us assume that our variation of the AF vanishes under certain circumstances: $\delta S[x] = 0$. We then say that the action, which remains

unchanged, is invariant under that particular variation of the path. Recall that the principle of stationary action then states:

$$\delta S[x] = 0 = G_2 - G_1,$$

i.e., G has the same value, independent of the initial and final configurations.

In particular, let us assume that the AF (in Hamiltonian formulation) is invariant for a variation around the actual path for which it holds that

$$\delta x^i(t_{1,2}) = 0, \qquad \frac{d}{dt}(\delta t) = 0, \qquad \text{therefore} \qquad \delta t = \text{const} = \varepsilon.$$

Then it follows from the invariance of the AF under infinitesimal constant time translation:

$$\delta S = 0 = G_2 - G_1 = -H(t_2)\delta t_2 + H(t_1)\delta t_1 = -(H_2 - H_1)\varepsilon,$$

the conservation of energy:

$$H(t_2) = H(t_1), \qquad \text{meaning} \qquad \dot{H} = 0.$$

Similarly, the conservation law for linear momentum follows if we assume that the AF is invariant under constant space translation and the change of the terminal times vanishes:

$$\begin{aligned} &\delta x^i = \delta \varepsilon_i = \text{const}, \qquad \delta t(t_{1,2}) = 0, \\ &\delta S = 0 = G_2 - G_1 = (p_i \delta x^i)_2 - (p_i \delta x^i)_1 = (p_{i_2} - p_{i_1})\delta \varepsilon_i, \\ \text{or} \qquad &p_i(t_2) = p_i(t_1), \qquad \text{meaning} \qquad \dot{p}_i = 0. \end{aligned}$$

Now let

$$H = \frac{p_i^2}{2m} + V(r),$$

i.e., potential may only depend on the distance $r = \sqrt{(x^i)^2}$. Then no space direction is distinguished, and with respect to rigid rotations $\delta\omega_i = \text{const}$ and

$$\delta t(t_{1,2}) = 0, \qquad \delta x^i = \varepsilon^{ij}_k \delta\omega_j x^k,$$

it can be proved that

$$\delta S = \delta \int_{t_1}^{t_2} dt \left[p_i \dot{x}^i - \frac{p_i^2}{2m} - V(\sqrt{x^i}) \right] = 0.$$

Because

$$\begin{aligned} \delta S = 0 = G_2 - G_1 &= (p_i \delta x^i)_2 - (p_i \delta x^i)_1 \\ &= (p_i \varepsilon^{ij}_k \delta\omega_j x^k)_2 - (p_i \varepsilon^{ij}_k \delta\omega_j x^k)_1 \\ &= \delta\omega_i \{ [(r \times p)_i]_2 - [(r \times p)_i]_1 \} \end{aligned}$$

this implies the conservation for angular momentum:

$$L(t_2) = L(t_1), \qquad \text{meaning} \qquad \dot{L} = 0.$$

Conversely, the conservation for angular momentum corresponds to the invariance, $\delta S = 0$, under rigid rotation in space. The generalization of this statement is this: if a conservation law exists, then the AF is stationary with respect to the infinitesimal transformation of a corresponding variable. The converse of this statement is also true: If the AF is invariant with respect to an infinitesimal transformation, $\delta S = 0$, then the corresponding conservation is valid.

Emmy Noether was primarily an algebraists, but when she came to Göttingen in 1915, she was asked by David Hilbert for help in trying to understand the status of *energy conservation* in general relativity. As we have seen, the conservation of energy in classical physics is closely related to the time–invariance of physical laws, but in general relativity there is not necessarily a global time coordinate, so the classical invariance cannot be invoked to establish the conservation of energy. Nevertheless, if spacetime in the region of interest is regarded as asymptotically flat, it is possible to define a conserved energy. This important aspect of general relativity was greatly clarified by Noether's Theorem in 1918. Subsequently the theorem has found important applications in many branches of physics. For example, in quantum mechanics the phase of the wave function can be incremented without affecting any observables, and this *gauge symmetry* corresponds to the *conservation of electric charge*. Moreover, Noether's approach of identifying symmetries with conserved quantities forms the basis of the Standard Model of particle physics.

A proper mathematical exposition of the Noether's theorem is founded on the theory of Sophus Lie: groups, algebras, symmetries and general invariance (see section 2.4.1 below).

1.1.5 Symplectic Mechanics

In general mechanics on smooth manifolds, one first defines the configuration manifold Q of the system in consideration, and then proceeds either using Lagrangian formalism on the tangent bundle TQ or Hamiltonian formalism on the cotangent bundle T^*Q. In case of Hamiltonian formalism, T^*Q is called the (momentum) phase space, admitting a natural symplectic structure that is usually defined as follows (see [AMR88, Arn89, Put93, MR99]). Let Q be a smooth n–dimensional manifold and pick local coordinates $\{dq^1, ..., dq^n\}$. Then $\{dq^1, ..., dq^n\}$ defines a basis of the cotangent space T^*_qQ, and by writing the canonical one–form $\theta \in T^*_qQ$ as

$$\theta = p_i dq^i, \tag{1.21}$$

we get local coordinates $\{q^1, ..., q^n, p_1, ..., p_n\}$ on T^*Q (see Figure 1.2 for the basis of Euclidean $\mathbb{R}^3-$space). Now, define the canonical symplectic two–form form ω on T^*Q by

$$\omega = d\theta = dp_i \wedge dq^i. \tag{1.22}$$

This two–form ω is obviously independent of the choice of coordinates $\{q^1, ..., q^n \in Q\}$ and independent of the base point $\{q^1, ..., q^n, p_1, ..., p_n\} \in T^*_q Q$; therefore, it is locally constant, and so $d\omega = 0$ (see Figure 1.3 for the fundamental two–form in Euclidean $\mathbb{R}^3-$space).

Let (M, ω) be a symplectic manifold and $H \in C^\infty(M, \mathbb{R})$ a smooth real valued function on M. Hamiltonian vector field X_H, corresponding to the total energy function H, is the smooth vector field on M, determined by the condition

$$i_{X_H}\omega + dH = 0,$$

where i_{X_H} denotes the contraction (or inner product) of the vector field X_H and the symplectic form w. A triple (M, ω, H) is called a Hamiltonian mechanical system. Nondegeneracy of ω guarantees that X_H exists (see [Put93]).

Let $\{q^1, ..., q^n, p_1, ..., p_n\}$ be canonical coordinates on M, i.e., relation (1.22) is valid. Then in these coordinates Hamiltonian vector field X_H is defined by

$$X_H = \frac{\partial H}{\partial p_i}\frac{\partial}{\partial q^i} - \frac{\partial H}{\partial q^i}\frac{\partial}{\partial p_i}. \tag{1.23}$$

As a consequence, $\big((q^i(t)), (p_i(t))\big)$ is an integral curve of X_H (for $i = 1, ..., n$) iff Hamilton's equations (3.10) hold.

In this way, the Newton's law of motion (1.1), for n conservative particles, has the following symplectic formulation on $\mathbb{R}^{3n}$ (see [AMR88, Put93]):

$$M = T^*\mathbb{R}^{3n} \simeq \mathbb{R}^{6n}, \qquad \omega = dp_i \wedge dq^i, \qquad H = \sum_{i=1}^{3n} \frac{p_i^2}{2m_i} + U.$$

The Hamiltonian vector field (1.23) is

$$X_H = \frac{p_i}{m_i}\frac{\partial}{\partial q^i} - \frac{\partial U}{\partial q^i}\frac{\partial}{\partial p_i},$$

and the Hamilton's equations (3.10) become

$$\dot{q}^i = \frac{p_i}{m_i}, \qquad \dot{p}_i = -\frac{\partial U}{\partial q^i}, \qquad (i = 1, \ldots, n).$$

1.1.6 Modern Rotational Biomechanics

Despite the elegance of translational symplectic geometry/mechanics outlined above, the most suitable formalism to deal with the full complexity of modern biomechanics of human motion and its robotics application, is *rotational symplectic geometry/mechanics*. Namely, as human joints are by nature rotational, combined muscular force vectors are transformed by joint geometry into *driving torque one–forms*, $T_i(t, q^i, p_i)$ (see Figure 1.4).

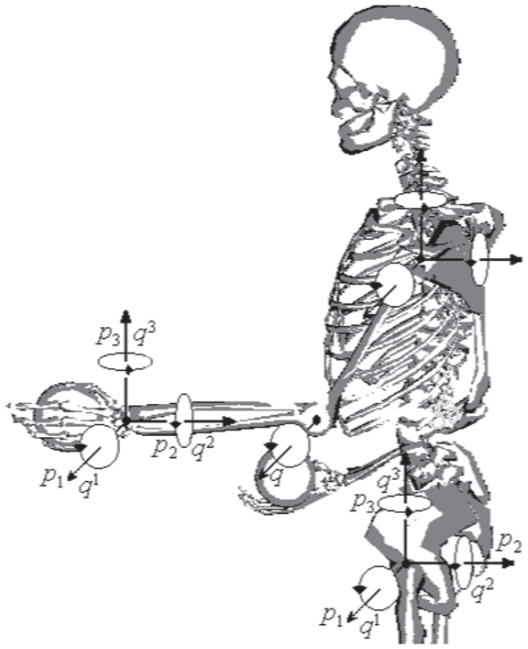


Fig. 1.4. Rotational synovial joints in human body.

Joint rotations are in modern biomechanics described in the language of constrained rotational Lie groups, $SO(2)$ and $SO(3)$, where constrained rotation is for each human joint separately defined in anatomically terms, so that the rotational joint amplitude is always less then the full circle. Nevertheless, the formalism of rotational Lie groups still works, just bearing in mind the imposed anatomical joint restrictions (see [Iva04, ILI95]).

All active joint–angles q^i $(i = 1, \ldots, N \equiv DOF)$, constitute a smooth configuration manifold Q^N, defined as a direct product of constrained rotational joint Lie groups $SO(3) \times SO(2) \times SO(3) \times ...$ for all rotational joints considered (see Figure 1.5). Uniaxial, 'hinge' joints represent constrained, classical, rotational groups $SO(2)^i$, parameterized by constrained angles $q^i \equiv q^i \in [q^i_{\min}, q^i_{\max}]$. Their associated velocities are defined by the corresponding Lie algebras $\mathfrak{so}(2)^i$. Three–axial, 'ball–and–socket' joints represent constrained rotational groups $SO(3)^i$, usually parameterized by constrained Euler angles $q^i_{1,2,3} = \{\phi, \psi, \theta\}^i$. Their associated velocities are defined by the corresponding Lie algebras $\mathfrak{so}(3)^i$.

We refer to the tangent bundle TQ^N of the configuration manifold Q^N as the *velocity phase–space* manifold, and to its cotangent bundle T^*Q^N as the *momentum phase–space* manifold. In this way, rotational biomechanics uses the full power of symplectic mechanics outlined above.

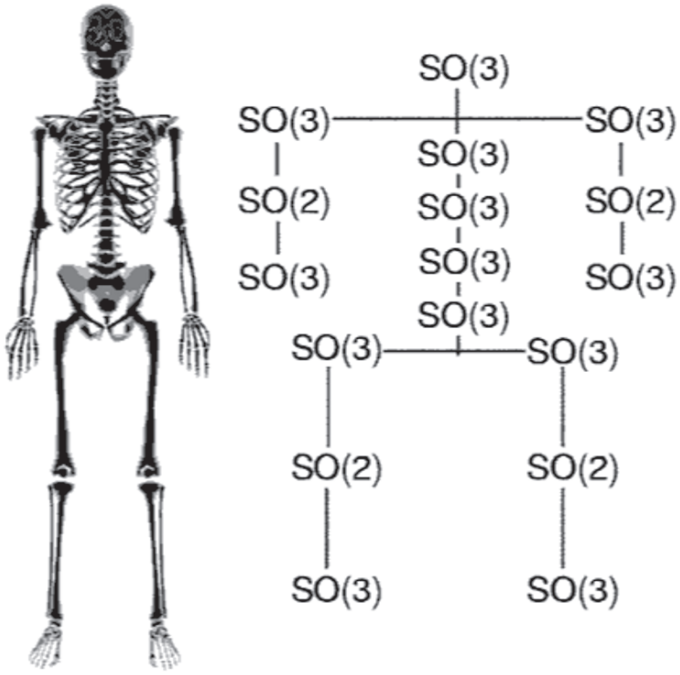


Fig. 1.5. Basic structure of the biomechanical configuration manifold Q^n composed as a direct product of constrained rotational Lie groups of human 'ball–end–socket joints'.

Passive Joint Dynamics

Recall that all biological systems are *dissipative structures*, emphasizing *irreversible processes* inefficient energetically, but highly efficient in terms of information and control (see [NP77]). In case of biomechanics, we have the passive joint damping contribution to *driving torque one–forms*, $T_i(t,\, q^i,\, p_i)$, which has the basic stabilizing effect to the complex human movement. This effect can be described by (q, p)–quadratic form of the *Rayleigh – Van der Pol's dissipation function* (see [BR78])

$$R = \frac{1}{2}\sum_{i=1}^{9} p_i^2\,[a_i \,+\, b_i(q^i)^2], \tag{1.24}$$

where a_i and b_i denote dissipation parameters. Its partial derivatives $\partial R/\partial p$ give rise to viscous forces in the joints which are linear in p_i and quadratic in q^i. It is based on the unforced Van der Pol's oscillator

$$\ddot{x} - \left(a + b\,x^2\right)\,\dot{x} + x = 0,$$

where the damping force $F^{dmp}(\dot{x}) = -\partial R/\partial \dot{x}$ is given by the Rayleigh's dissipation function $R = \frac{1}{2}\left(a + b\,x^2\right)\dot{x}^2$ – with the velocity term $\dot{x}$ replaced by our momentum term p^2 (see [Iva04, ILI95]).

Using (1.24) we get the *dissipative Hamiltonian biomechanics*

$$\begin{aligned} \dot{q}^i &= \frac{\partial H(q,p)}{\partial p_i} + \frac{\partial R(q,p)}{\partial p_i}, \\ \dot{p}_i &= -\frac{\partial H(q,p)}{\partial q^i} + \frac{\partial R(q,p)}{\partial q^i}, \qquad (i = 1, \ldots, N), \end{aligned} \tag{1.25}$$

which reduces to the gradient system in case $H = 0$ (as well as to the conservative system in case $R = 0$).

1.1.7 Muscular Dynamics and Control

Muscular Dynamics

Muscular dynamics describes the internal *excitation* and *contraction* dynamics [Hat78, Iva04] of *equivalent muscular actuators*, anatomically represented by resulting action of *antagonistic muscle–pairs* for each uniaxial joint. We attempt herein to describe the equivalent muscular dynamics in the simplest possible way (for example, Hatze used 51 nonlinear differential equations of the first order to derive his, arguably most elaborate, myocybernetic model [Hat78]), and yet to include the main excitation and contraction relations.

The active muscular–control contribution to the torque one–forms, $T_i = T_i(t,\, q^i,\, p_i)$, should describe the internal *excitation* and *contraction* dynamics [IS01, Iva04, Iva91, IP01a]) of *equivalent muscular actuators*, anatomically represented by resulting action of *antagonistic muscle–pairs* per each active degree–of–freedom.

(a) *Excitation dynamics* can be described by impulse *torque–time* relation

$$\begin{aligned} T_i^{imp} &= T_i^0(1 - e^{-t/\tau_i}) \qquad \text{if stimulation } > 0 \\ T_i^{imp} &= T_i^0 e^{-t/\tau_i} \qquad\quad \text{if stimulation } = 0, \end{aligned}$$

where F_i^0 denote the maximal isometric muscular torques applied at i–th joint, while τ_i denote the time characteristics of particular muscular actuators. This is a rotational–joint form of the solution of the Wilkie's *muscular active–state element equation* [Wil56]

$$\dot{x} + \beta\, x = \beta\, S\, A, \quad x(0) = 0, \quad 0 < S < 1,$$

where $x = x(t)$ represents the active state of the muscle, β denotes the element gain, A corresponds to the maximum tension the element can develop, and $S = S(r)$ is the 'desired' active state as a function of motor unit stimulus rate r.

(b) *Contraction dynamics* has classically been described by the Hill's *hyperbolic force–velocity* relation [Hil38], which we propose here in the rotational (q,p)–form

$$T_i^{Hill} = \frac{\left(T_i^0 b_i - a_i p_i\right)}{(p_i - b_i)},$$

where a_i (having dimension of torque) and b_i (having dimension of momentum) denote the *rotational Hill's parameters* (see [IS01, Iva04]), corresponding to the energy dissipated during the contraction and the phosphagenic energy conversion rate, respectively.

Therefore, we can describe the excitation/contraction dynamics for the ith equivalent muscle–joint actuator, i.e., antagonistic muscle pair, by the simple impulse–hyperbolic product–relation

$$T_i(t,q,p) = T_i^{imp} \times T_i^{Hill}, \qquad (i = 1,\dots,N). \tag{1.26}$$

Using (1.26) we get the *forced dissipative Hamiltonian biomechanics*, in the form

$$\begin{aligned} \dot{q}^i &= \frac{\partial H(q,p)}{\partial p_i} + \frac{\partial R(q,p)}{\partial p_i}, \qquad (i = 1,\dots,N), \\ \dot{p}_i &= T_i(t,q,p) - \frac{\partial H(q,p)}{\partial q^i} + \frac{\partial R(q,p)}{\partial q^i}. \end{aligned} \tag{1.27}$$

Muscular Control

We introduce the *control Hamiltonian function* $H_c : T^*M^N \to \mathbb{R}$, in local canonical coordinates on T^*M^N defined by [NS90]

$$H_c(q,p,u) = H_0(q,p) - q^i\, u_i, \qquad (i = 1,\dots,N), \tag{1.28}$$

where $u_i = u_i(t,q,p)$ are feedback–control one–forms, representing the corrections to the torque one–forms $T_i(t,q,p)$.

Using (1.28), the affine Hamiltonian system can be defined as

$$\begin{aligned} \dot{q}^i &= \frac{\partial H_c(q,p,u)}{\partial p_i} + \frac{\partial R(q,p)}{\partial p_i}, \qquad (i = 1,\dots,N) \\ \dot{p}_i &= T_i(t,q,p) - \frac{\partial H_c(q,p,u)}{\partial q^i} + \frac{\partial R(q,p)}{\partial q^i}, \\ o^i &= -\frac{\partial H_c(q,p,u)}{\partial u_i}, \qquad q^i(0) = q_0^i, \qquad p_i(0) = p_i^0, \end{aligned} \tag{1.29}$$

where $o^i = o^i(t)$ represent the natural outputs which can be different from commonly used joint angles.

If nominal reference outputs $o_R^i = o_R^i(t)$ are known, the simple PD *stiffness–servo* [Whi87] could be formulated, via *error function* $e(t) = o^j - o_R^j$, in covariant form

$$u_i = K_o \delta_{ij}(o^j - o^j_R) + K_{\dot{o}} \delta_{ij}(\dot{o}^j - \dot{o}^j_R), \tag{1.30}$$

where Ks are the control–gains and δ_{ij} is the Kronecker tensor.

If natural outputs o^i *actually are* the joint angles and nominal canonical trajectories $\left(q^i_R = q^i_R(t),\, p^R_i = p^R_i(t)\right)$ are known, then the stiffness–servo (1.30) could be formulated in canonical form as

$$u_i = K_q \delta_{ij}(q^i - q^i_R) + K_p(p_i - p^R_i). \tag{1.31}$$

In this way formulated affine Hamiltonian control system (1.29–1.31) resembles the physiological *autogenetic motor servo* [Hou79], acting on the spinal–reflex level of the human locomotor control in the following way. Voluntary contraction force Φ of human skeletal muscle is reflexly excited (positive feedback $+\Phi^{-1}$) by responses of its *spindle receptors* to stretch and is reflexly inhibited (negative feedback $-\Phi^{-1}$) by responses of its *Golgi tendon organs* to contraction. Stretch and unloading reflexes are mediated by combined actions of several autogenetic neural pathways, forming the so-called 'motor–servo'. Term '*autogenetic*' means that the stimulus excites receptors located in the same muscle that is the target of the reflex response. The most important of these muscle receptors are the primary and secondary endings in muscle-spindles, sensitive to length change – positive length feedback $+\Phi^{-1}$, and the Golgi tendon organs, sensitive to contractile force – negative force feedback $-\Phi^{-1}$.

The gain G of the length feedback $+\Phi^{-1}$ can be expressed as the *positional stiffness* (the ratio $G \approx S = d\Phi/dx$ of the force–Φ change to the length–x change) of the muscle system. The greater the stiffness S, the less will the muscle be disturbed by a change in load and the more reliable will be the performance of the muscle system in executing controlled changes in length $+\Phi^{-1}$.

The autogenetic circuits $+\Phi^{-1}$ and $-\Phi^{-1}$ appear to function as *servoregulatory loops* that convey continuously graded amounts of excitation and inhibition to the large (*alpha*) skeletomotor neurons. Small (*gamma*) fusimotor neurons innervate the contractile poles of muscle spindles and function to modulate spindle–receptor discharge (for further details, see section 5.4 below).

1.2 Global Functorial Language of Human–Like Biomechanics

In modern mathematical sciences whenever one defines a new class of mathematical objects, one proceeds almost in the next breath to say what kinds of maps between objects will be considered [Swi75]. A general framework for dealing with situations where we have some *objects* and *maps between objects*, like sets and functions, vector spaces and linear operators, points in a

space and paths between points, etc. – provides the modern metalanguage of categories and functors. Categories are mathematical universes and functors are 'projectors' from one universe onto another. For this reason, in this book we extensively use this language, mainly following its founder, S. MacLane [MacL71].

1.2.1 Preliminaries from Calculus, Algebra and Topology

Before defining categories, functors and their natural transformations, we give the necessary preliminaries from calculus, algebra and point–set topology.

Notes From Calculus

Functions

Recall that a *function* f is a *rule* that assigns to each element x in a set A exactly one element, called $f(x)$, in a set B. A function could be thought of as a *machine* $[[f]]$ with $x-$input (the *domain* of f is the set of all possible inputs) and $f(x)-$output (the *range* of f is the set of all possible outputs) [Stu99]

$$x \to [[f]] \to f(x)$$

There are four possible ways to represent a function: (i) verbally (by a description in words); (ii) numerically (by a table of values); (iii) visually (by a graph); and (iv) algebraically (by an explicit formula). The most common method for visualizing a function is its graph. If f is a function with domain A, then its *graph* is the set of ordered input–output pairs

$$\{(x, f(x)) : x \in A\}.$$

Algebra of Functions

Let f and g be functions with domains A and B. Then the functions $f+g$, $f-g$, fg, and f/g are defined as follows [Stu99]

$$\begin{aligned}
(f+g)(x) &= f(x)+g(x) && \text{domain } = A \cap B,\\
(f-g)(x) &= f(x)-g(x) && \text{domain } = A \cap B,\\
(fg)(x) &= f(x)\,g(x) && \text{domain } = A \cap B,\\
\left(\frac{f}{g}\right)(x) &= \frac{f(x)}{g(x)} && \text{domain } = \{x \in A \cap B : g(x) \neq 0\}.
\end{aligned}$$

Compositions of Functions

Given two functions f and g, the composite function $f \circ g$ (also called the *composition* of f and g) is defined by

$$(f \circ g)(x) = f(g(x)).$$

The $(f \circ g)-$machine is composed of the $g-$machine (first) and then the $f-$machine [Stu99],

$$x \to [[g]] \to g(x) \to [[f]] \to f(g(x))$$

For example, suppose that $y = f(u) = \sqrt{u}$ and $u = g(x) = x^2 + 1$. Since y is a function of u and u is a function of x, it follows that y is ultimately a function of x. We compute this by substitution

$$y = f(u) = f \circ g = f(g(x)) = f(x^2 + 1) = \sqrt{x^2 + 1}.$$

The Chain Rule

If f and g are both differentiable and $h = f \circ g$ is the composite function defined by $h(x) = f(g(x))$, then h is differentiable and h' is given by the product [Stu99]

$$h'(x) = f'(g(x))\, g'(x).$$

In Leibniz notation, if $y = f(u)$ and $u = g(x)$ are both differentiable functions, then

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}.$$

The reason for the name *chain rule* becomes clear if we add another link to the chain. Suppose that we have one more differentiable function $x = h(t)$. Then, to compute the derivative of y with respect to t, we use the chain rule twice,

$$\frac{dy}{dt} = \frac{dy}{du}\frac{du}{dx}\frac{dx}{dt}.$$

Integration and Change of Variables

Based on the chain rule, under the certain hypotheses (such as a one–to–one C^0 map T with a nonzero *Jacobian* $\left|\frac{\partial(x,\ldots)}{\partial(u,\ldots)}\right|$ that maps a region S onto a region R, see [Stu99]) we have the following substitution formulas:

1. for a single integral,

$$\int_R f(x)\, dx = \int_S f(x(u))\frac{\partial x}{\partial u} du,$$

2. for a double integral,

$$\iint_R f(x,y)\,dA = \iint_S f(x(u,v),y(u,v))\left|\frac{\partial(x,y)}{\partial(u,v)}\right| dudv,$$

3. for a triple integral,

$$\iiint_R f(x,y,z)\,dV = \iint_S f(x(u,v,w),y(u,v,w),z(u,v,w))\left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right| dudvdw,$$

4. similarly for $n-$tuple integrals.

Notes from Set Theory

Given a function $f : A \to B$, the set A is called the *domain* of f, and denoted $\mathrm{Dom}\, f$. The set B is called the *codomain* of f, and denoted $\mathrm{Cod}\, f$. The codomain is not to be confused with the *range* of $f(A)$, which is in general only a subset of B.

A function $f : X \to Y$ is called *injective* or *one–to–one* or an *injection* if for every y in the codomain Y there is at most one x in the domain X with $f(x) = y$. Put another way, given x and x' in X, if $f(x) = f(x')$, then it follows that $x = x'$. A function $f : X \to Y$ is called *surjective* or *onto* or a *surjection* if for every y in the codomain $\mathrm{Cod}\, f$ there is at least one x in the *domain* X with $f(x) = y$. Put another way, the *range* $f(X)$ is equal to the codomain Y. A function is *bijective* iff it is both injective and surjective. Injective functions are called the *monomorphisms*, and surjective functions are called the *epimorphisms* in the *category of sets* (see below).

A *relation* is any subset of a *Cartesian product* (see below). By definition, an *equivalence relation* α on a set X is a relation which is *reflexive, symmetrical* and *transitive*, i.e., relation that satisfies the following three conditions:

1. *Reflexivity*: each element $x \in X$ is equivalent to itself, i.e., $x\alpha x$,
2. *Symmetry*: for any two elements $x, x' \in X$, $x\alpha x'$ implies $x'\alpha x$, and
3. *Transitivity*: $a \leq b$ and $b \leq c$ implies $a \leq c$.

Similarly, a relation $\leq$ defines a *partial order* on a set S if it has the following properties:

1. *Reflexivity*: $a \leq a$ for all $a \in S$,
2. *Antisymmetry*: $a \leq b$ and $b \leq a$ implies $a = b$, and
3. *Transitivity*: $a \leq b$ and $b \leq c$ implies $a \leq c$.

A *partially ordered set* (or *poset*) is a set taken together with a partial order on it. Formally, a partially ordered set is defined as an ordered pair $P = (X, \leq)$, where X is called the *ground set* of P and $\leq$ is the partial order of P.

Notes from General Topology

Topology is a kind of *abstraction* of Euclidean geometry, and also a natural framework for the study of *continuity*. Euclidean geometry is abstracted by regarding triangles, circles, and squares as being the same basic object. Continuity enters because in saying this one has in mind a *continuous deformation* of a triangle into a square or a circle, or any arbitrary shape. On the other hand, a disk with a hole in the center is topologically different from a circle or a square because one cannot create or destroy holes by continuous deformations. Thus using topological methods one does not expect to be able to identify a geometric figure as being a triangle or a square. However, one does expect to be able to detect the presence of gross features such as holes or the fact that the figure is made up of two disjoint pieces etc. In this way topology produces theorems that are usually qualitative in nature – they may assert, for example, the existence or non–existence of an object. They will not in general, provide the means for its construction [Nas83].

Let X be any *set* and $Y = \{X_\alpha\}$ denote a collection, finite or infinite of subsets of X. Then X and Y form a *topological space* provided the X_α and Y satisfy:

1. Any finite or infinite subcollection $\{Z_\alpha\} \subset X_\alpha$ has the property that $\cup Z_\alpha \in Y$, and
2. Any *finite subcollection* $\{Z_{\alpha_1}, ..., Z_{\alpha_n}\} \subset X_\alpha$ has the property that $\cap Z_{\alpha_i} \in Y$.

The set X is then called a topological space and the X_α are called *open sets*. The choice of Y satisfying (2) is said to give a topology to X.

Now, given two topological spaces X and Y, a *function* (or, a *map*) $f : X \to Y$ is *continuous* if the inverse image of an open set in Y is an open set in X.

The main general idea in topology is to study spaces which can be continuously deformed into one another, namely the idea of *homeomorphism*. If we have two topological spaces X and Y, then a map $f : X \to Y$ is called a homeomorphism iff

1. f is continuous, and
2. There exists an inverse of f, denoted f^{-1}, which is also continuous.

Definition (2) implies that if f is a homeomorphism then so is f^{-1}. Homeomorphism is the main topological example of *reflexive*, *symmetrical* and *transitive relation*, i.e., *equivalence relation*. Homeomorphism divides all topological spaces up into *equivalence classes*. In other words, a pair of topological spaces, X and Y, belong to the same equivalence class if they are homeomorphic.

The second example of topological equivalence relation is *homotopy*. While homeomorphism generates equivalence classes whose members are topological spaces, homotopy generates equivalence classes whose members are continuous maps (or, C^0-maps). Consider two continuous maps $f, g : X \to Y$ between

topological spaces X and Y. Then the map f is said to be *homotopic* to the map g if f can be continuously deformed into g (see below for the precise definition of homotopy). Homotopy is an equivalence relation which divides the space of continuous maps between two topological spaces into equivalence classes [Nas83].

Another important notions in topology are *covering*, *compactness* and *connectedness*. Given a family of sets $\{X_\alpha\} = X$ say, then X is a *cover* of another set Y if $\cup X_\alpha$ contains Y. If all the X_α happen to be open sets the cover is called an *open cover*. Now consider the set Y and all its possible open coverings. The set Y is *compact* if for every open covering $\{X_\alpha\}$ with $\cup X_\alpha \supset Y$ there always exists a finite subcovering $\{X_1, ..., X_n\}$ of Y with $X_1 \cup ... \cup X_n \supset Y$.[5] Again, we define a set Z to be *connected* if it cannot be written as $Z = Z_1 \cup Z_2$, where Z_1 and Z_2 are both open and $Z_1 \cap Z_2$ is an empty set.

Let $A_1, A_2, ..., A_n$ be closed subspaces of a topological space X such that $X = \cup_{i=1}^n A_i$. Suppose $f_i : A_i \to Y$ is a function, $1 \leq i \leq n$, iff

$$f_i | A_i \cap A_j = f_j | A_i \cap A_j, 1 \leq i, j \leq n. \tag{1.32}$$

In this case f is continuous iff each f_i is. Using this procedure we can define a C^0-function $f : X \to Y$ by cutting up the space X into closed subsets A_i and defining f on each A_i separately in such a way that $f|A_i$ is obviously continuous; we then have only to check that the different definitions agree on the *overlaps* $A_i \cap A_j$.

The *universal property of the Cartesian product*: let $p_X : X \times Y \to X$, and $p_Y : X \times Y \to Y$ be the *projections* onto the first and second factors, respectively. Given any pair of functions $f : Z \to X$ and $g : Z \to Y$ there is a unique function $h : Z \to X \times Y$ such that $p_X \circ h = f$, and $p_Y \circ h = g$. Function h is continuous iff both f and g are. This property characterizes X/α up to homeomorphism. In particular, to check that a given function $h : Z \to X$ is continuous it will suffice to check that $p_X \circ h$ and $p_Y \circ h$ are continuous.

The *universal property of the quotient*: let α be an equivalence relation on a topological space X, let X/α denote the *space of equivalence classes* and $p_\alpha : X \to X/\alpha$ the *natural projection*. Given a function $f : X \to Y$, there is a function $f' : X/\alpha \to Y$ with $f' \circ p_\alpha = f$ iff $x \alpha x'$ implies $f(x) = f(x')$, for all $x \in X$. In this case f' is continuous iff f is. This property characterizes X/α up to homeomorphism.

Now we return to the fundamental notion of homotopy. Let I be a compact unit interval $I = [0, 1]$. A *homotopy* from X to Y is a continuous function

[5] The notion of compactness is fundamental for biomechanical control. Namely, the basic (kinematic) unit of the biomechanical manifold is the special Euclidean group $SE(3)$. This group is non–compact, which means that it does not admit a natural metric generated by the segment's kinetic energy, and therefore there is not a natural control. However, its two subgroups, the group of rotations $SE(3)$ and the group of translations $\mathbb{R}^3$ are both compact, admitting the natural quadratic metric forms given by the kinetic energy. This implies the existence of (muscular–like) optimal controls in the sense of geodesics (see Chapter 2).

$F : X \times I \to Y$. For each $t \in I$ one has $F_t : X \to Y$ defined by $F_t(x) = F(x,t)$ for all $x \in X$. The functions F_t are called the 'stages' of the homotopy. If $f, g : X \to Y$ are two continuous maps, we say f is homotopic to g, and write $f \simeq g$, if there is a homotopy $F : X \times I \to Y$ such that $F_0 = f$ and $F_1 = g$. In other words, f can be continuously deformed into g through the stages F_t. If $A \subset X$ is a subspace, then F is a homotopy relative to A if $F(a,t) = F(a,0)$, for all $a \in A, t \in I$.

The homotopy relation $\simeq$ is an equivalence relation. To prove that we have $f \simeq f$ is obvious; take $F(x, t = f(x)$, for all $x \in X$, $t \in I$. If $f \simeq g$ and F is a homotopy from f to g, then $G : X \times I \to Y$ defined by $G(x,t) = F(x, 1-t)$, is a homotopy from g to f, i.e., $g \simeq f$. If $f \simeq g$ with homotopy F and $g \simeq f$ with homotopy G, then $f \simeq h$ with homotopy H defined by

$$H(x,t) = \begin{cases} F(x,t), & 0 \leq t \leq 1/2 \\ G(x, 2t-1), & 1/2 \leq t \leq 1 \end{cases} .$$

To show that H is continuous we use the relation (1.32).

In this way, the set of all C^0-functions $f : X \to Y$ between two topological spaces X and Y, called the *function space* and denoted by Y^X, is partitioned into equivalence classes under the relation $\simeq$. The equivalence classes are called *homotopy classes*, the homotopy class of f is denoted by $[f]$, and the set of all homotopy classes is denoted by $[X;Y]$.

If α is an equivalence relation on a topological space X and $F : X \times I \to Y$ is a homotopy such that each stage F_t factors through X/α, i.e., $x \alpha x'$ implies $F_t(x) = F_t(x')$, then F induces a homotopy $F' : (X/\alpha) \times I \to Y$ such that $F' \circ (p_\alpha \times 1) = F$.

Homotopy theory has a range of applications of its own, outside topology and geometry, as for example in proving Cauchy theorem in complex variable theory, or in solving nonlinear equations of artificial neural networks.

A *pointed set* (S, s_0) is a set S together with a distinguished point $s_0 \in S$. Similarly, a *pointed topological space* (X, x_0) is a space X together with a distinguished point $x_0 \in X$. When we are concerned with pointed spaces $(X, x_0), (Y, y_0)$, etc., we always require that all functions $f : X \to Y$ shell preserve base points, i.e., $f(x_0) = y_0$, and that all homotopies $F : X \times I \to Y$ be relative to the base point, i.e., $F(x_0, t) = y_0$, for all $t \in I$. We denote the homotopy classes of base point–preserving functions by $[X, x_0; Y, y_0]$ (where homotopies are relative to x_0). $[X, x_0; Y, y_0]$ is a pointed set with base point f_0, the constant function: $f_0(x) = y_0$, for all $x \in X$.

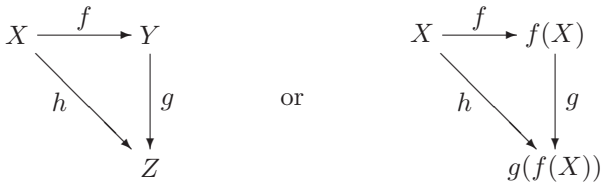
A *path* $\gamma(t)$ from x_0 to x_1 in a topological space X is a continuous map $\gamma : I \to X$ with $\gamma(0) = x_0$ and $\gamma(1) = x_1$. Thus X^I is the space of all paths in X with the compact–open topology. We introduce a relation $\sim$ on X by saying $x_0 \sim x_1$ iff there is a path $\gamma : I \to X$ from x_0 to x_1. $\sim$ is clearly an equivalence relation, and the set of equivalence classes is denoted by $\pi_0(X)$. The elements of $\pi_0(X)$ are called the *path components*, or $0-$*components* of X. If $\pi_0(X)$ contains just one element, then X is called *path connected*, or $0-$*connected*. A *closed path*, or *loop* in X at the point x_0 is a path $\gamma(t)$ for

which $\gamma(0) = \gamma(1) = x_0$. The *inverse loop* $\gamma^{-1}(t)$ based at $x_0 \in X$ is defined by $\gamma^{-1}(t) = \gamma(1-t)$, for $0 \leq t \leq 1$. The *homotopy of loops* is the particular case of the above defined homotopy of continuous maps.

If (X, x_0) is a pointed space, then we may regard $\pi_0(X)$ as a pointed set with the 0−component of x_0 as a base point. We use the notation $\pi_0(X, x_0)$ to denote $p_0(X, x_0)$ thought of as a pointed set. If $f : X \to Y$ is a map then f sends 0−components of X into 0−components of Y and hence defines a function $\pi_0(f) : \pi_0(X) \to \pi_0(Y)$. Similarly, a base point–preserving map $f : (X, x_0) \to (Y, y_0)$ induces a map of pointed sets $\pi_0(f) : \pi_0(X, x_0) \to \pi_0(Y, y_0)$. In this way defined π_0 represents a 'functor' from the 'category' of topological (point) spaces to the underlying category of (point) sets (see the next section).

Commutative Diagrams

S. MacLane says that the category theory was born with an observation that many properties of mathematical systems can be unified and simplified by a presentation with *commutative diagrams of arrows* [MacL71]. Each arrow $f : X \to Y$ represents a function (i.e., a map, transformation, operator); that is, a source (domain) set X, a target (codomain) set Y, and a rule $x \mapsto f(x)$ which assigns to each element $x \in X$ an element $f(x) \in Y$. A typical diagram of sets and functions is



This diagram is *commutative* iff $h = g \circ f$, where $g \circ f$ is the usual composite function $g \circ f : X \to Z$, defined by $x \mapsto g(f(x))$.

Similar commutative diagrams apply in other mathematical, physical and computing contexts; e.g., in the 'category' of all topological spaces, the letters X, Y, and Z represent topological spaces while f, g, and h stand for continuous maps. Again, in the category of all groups, X, Y, and Z stand for groups, f, g, and h for homomorphisms.

Less formally, composing maps is like following directed paths from one object to another (e.g., from set to set). In general, a diagram is commutative iff any two paths along arrows that start at the same point and finish at the same point yield the same 'homomorphism' via compositions along successive arrows. Commutativity of the whole diagram follows from commutativity of its triangular components (depicting a 'commutative flow', see Figure 1.6). Study of commutative diagrams is popularly called 'diagram chasing', and provides a powerful tool for mathematical thought.

As an example from linear algebra, consider an elementary diagrammatic description of matrices, using the following *pull–back diagram* [Bar93]:

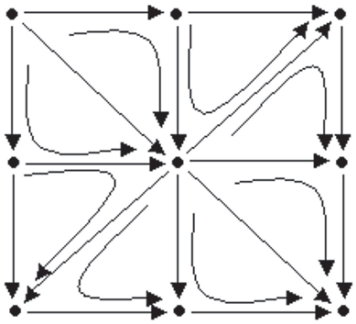
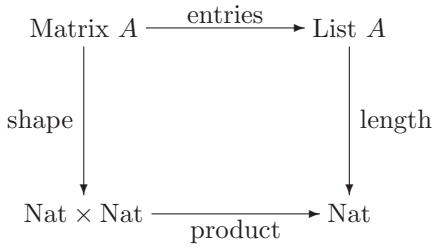
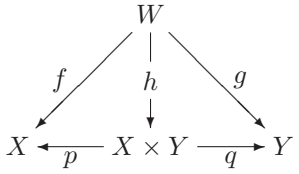


Fig. 1.6. A *commutative flow* (denoted by curved arrows) on a *triangulated digraph*. Commutativity of the whole diagram follows from commutativity of its triangular components.



asserts that a matrix is determined by its shape, given by a pair of natural numbers representing the number of rows and columns, and its data, given by the matrix entries listed in some specified order.

Many properties of mathematical constructions may be represented by *universal properties* of diagrams [MacL71]. Consider the *Cartesian product* $X \times Y$ of two sets, consisting as usual of all ordered pairs $\langle x, y \rangle$ of elements $x \in X$ and $y \in Y$. The projections $\langle x, y \rangle \mapsto x$, $\langle x, y \rangle \mapsto y$ of the product on its 'axes' X and Y are functions $p : X \times Y \to X$, $q : X \times Y \to Y$. Any function $h : W \to X \times Y$ from a third set W is uniquely determined by its composites $p \circ h$ and $q \circ h$. Conversely, given W and two functions f and g as in the diagram below, there is a unique function h which makes the following diagram commute:



This property describes the Cartesian product $X \times Y$ uniquely; the same diagram, read in the category of topological spaces or of groups, describes uniquely the Cartesian product of spaces or of the direct product of groups.

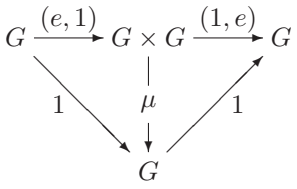
The construction 'Cartesian product' is technically called a 'functor' because it applies suitably both to the sets and to the functions between them; two functions $k : X \to X'$ and $l : Y \to Y'$ have a function $k \times l$ as their Cartesian product:

$$k \times l : X \times Y \to X' \times Y', \qquad \langle x, y\rangle \mapsto \langle kx, ly\rangle.$$

Groups and Related Algebraic Structures

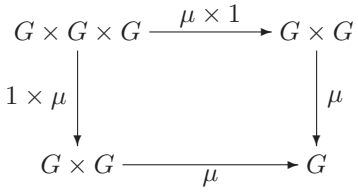
As already stated, the basic functional unit of lower biomechanics is the special Euclidean group $SE(3)$ of rigid body motions. In general, a *group* is a pointed set (G, e) with a *multiplication* $\mu : G \times G \to G$ and an *inverse* $\nu : G \to G$ such that the following diagrams commute [Swi75]:

1.



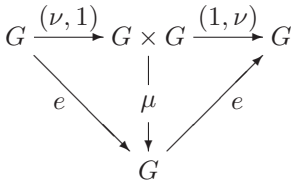
(e is a two–sided identity)

2.



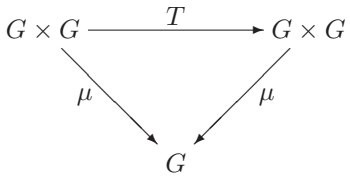
(associativity)

3.



(inverse).

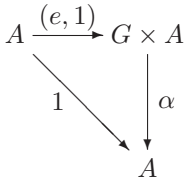
Here $e : G \to G$ is the constant map $e(g) = e$ for all $g \in G$. $(e, 1)$ means the map such that $(e, 1)(g) = (e, g)$, etc. A group G is called *commutative* or *Abelian group* if in addition the following diagram commutes



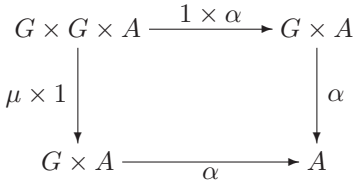
where $T : G \times G \to G \times G$ is the switch map $T(g_1, g_2) = (g_1, g_2)$, for all $(g_1, g_2) \in G \times G$.

A group G *acts* (on the left) on a set A if there is a function $\alpha : G \times A \to A$ such that the following diagrams commute [Swi75]:

1.



2.



where $(e, 1)(x) = (e, x)$ for all $x \in A$. The *orbits* of the action are the sets $Gx = \{gx : g \in G\}$ for all $x \in A$.

Given two groups $(G, *)$ and $(H, \cdot)$, a *group homomorphism* from $(G, *)$ to $(H, \cdot)$ is a function $h : G \to H$ such that for all x and y in G it holds that

$$h(x * y) = h(x) \cdot h(y).$$

From this property, one can deduce that h maps the identity element e_G of G to the identity element e_H of H, and it also maps inverses to inverses in the sense that $h(x^{-1}) = h(x)^{-1}$. Hence one can say that h is *compatible* with the *group structure*.

The *kernel* $\operatorname{Ker} h$ of a group homomorphism $h : G \to H$ consists of all those elements of G which are sent by h to the identity element e_H of H, i.e.,

$$\operatorname{Ker} h = \{x \in G : h(x) = e_H\}.$$

The *image* $\operatorname{Im} h$ of a group homomorphism $h : G \to H$ consists of all elements of G which are sent by h to H, i.e.,

$$\operatorname{Im} h = \{h(x) : x \in G\}.$$

The kernel is a *normal subgroup* of G and the image is a *subgroup* of H. The homomorphism h is *injective* (and called a *group monomorphism*) iff $\operatorname{Ker} h = e_G$, i.e., iff the kernel of h consists of the identity element of G only.

Similarly, a *ring* is a set S together with two binary operators $+$ and $*$ (commonly interpreted as addition and multiplication, respectively) satisfying the following conditions:

1. Additive associativity: For all $a, b, c \in S$, $(a + b) + c = a + (b + c)$,
2. Additive commutativity: For all $a, b \in S$, $a + b = b + a$,
3. Additive identity: There exists an element $0 \in S$ such that for all $a \in S$, $0 + a = a + 0 = a$,
4. Additive inverse: For every $a \in S$ there exists $-a \in S$ such that $a+(-a) = (-a) + a = 0$,
5. Multiplicative associativity: For all $a, b, c \in S$, $(a * b) * c = a * (b * c)$,
6. Left and right distributivity: For all $a, b, c \in S$, $a*(b+c) = (a*b)+(a*c)$ and $(b + c) * a = (b * a) + (c * a)$.

A ring (the term introduced by *D.Hilbert*) is therefore an Abelian group under addition and a semigroup under multiplication. A ring that is commutative under multiplication, has a unit element, and has no divisors of zero is called an *integral domain*. A ring which is also a commutative multiplication group is called a *field*. The simplest rings are the integers $\mathbb{Z}$, polynomials $\mathbb{R}[x]$ and $\mathbb{R}[x, y]$ in one and two variables, and square $n \times n$ real matrices.

An *ideal* is a subset $\mathfrak{I}$ of elements in a ring R which forms an additive group and has the property that, whenever x belongs to R and y belongs to $\mathfrak{I}$, then xy and yx belong to $\mathfrak{I}$. For example, the set of even integers is an ideal in the ring of integers $\mathbb{Z}$. Given an ideal $\mathfrak{I}$, it is possible to define a factor ring $R/\mathfrak{I}$.

A ring is called *left* (respectively, *right*) *Noetherian* if it does not contain an infinite ascending chain of left (respectively, right) ideals. In this case, the ring in question is said to satisfy the ascending chain condition on left (respectively, right) ideals. A *ring* is said to be *Noetherian* if it is both left and right Noetherian. If a ring R is Noetherian, then the following are equivalent:

1. R satisfies the ascending chain condition on ideals.
2. Every ideal of R is finitely generated.
3. Every set of ideals contains a maximal element.

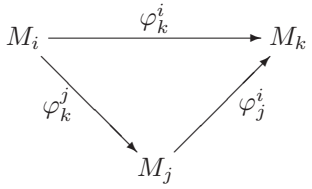
A *module* is a mathematical object in which things can be added together commutatively by multiplying coefficients and in which most of the rules of manipulating vectors hold. A module is abstractly very similar to a vector space, although in modules, coefficients are taken in rings which are much more general algebraic objects than the fields used in vector spaces. A module taking its coefficients in a ring R is called a module over R or $R-$module. Modules are the basic tool of homological algebra.

Examples of modules include the set of integers $\mathbb{Z}$, the cubic lattice in d dimensions $\mathbb{Z}^d$, and the group ring of a group. $\mathbb{Z}$ is a module over itself. It is closed under addition and subtraction. Numbers of the form $n\alpha$ for $n \in \mathbb{Z}$ and α a fixed integer form a submodule since, for $(n, m) \in \mathbb{Z}$, $n\alpha \pm m\alpha = (n \pm m)\alpha$ and $(n \pm m)$ is still in $\mathbb{Z}$. Also, given two integers a and b, the smallest module containing a and b is the module for their greatest common divisor, $\alpha = GCD(a, b)$.

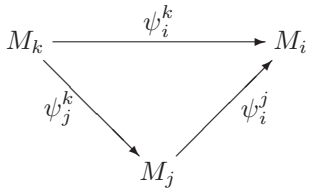
A module M is a *Noetherian module* if it obeys the ascending chain condition with respect to inclusion, i.e., if every set of increasing sequences of submodules eventually becomes constant. If a module M is Noetherian, then the following are equivalent:

1. M satisfies the ascending chain condition on submodules.
2. Every submodule of M is finitely generated.
3. Every set of submodules of M contains a maximal element.

Let I be a partially ordered set. A *direct system* of $R-$*modules* over I is an ordered pair $\{M_i, \varphi_j^i\}$ consisting of an indexed family of modules $\{M_i : i \in I\}$ together with a family of homomorphisms $\{\varphi_j^i : M_i \to M_j\}$ for $i \leq j$, such that $\varphi_i^i = 1_{M_i}$ for all i and such that the following diagram commutes whenever $i \leq j \leq k$



Similarly, an *inverse system* of $R-$*modules* over I is an ordered pair $\{M_i, \psi_i^j\}$ consisting of an indexed family of modules $\{M_i : i \in I\}$ together with a family of homomorphisms $\{\psi_i^j : M_j \to M_i\}$ for $i \leq j$, such that $\psi_i^i = 1_{M_i}$ for all i and such that the following diagram commutes whenever $i \leq j \leq k$



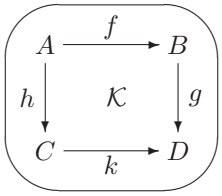
1.2.2 Categories

A category is a generic mathematical structure consisting of a collection of *objects* (sets with possibly additional structure), with a corresponding collection of *arrows*, or *morphisms*, between objects (agreeing with this additional structure). A category $\mathcal{K}$ is defined as a pair $(\mathtt{Ob}(\mathcal{K}), \mathtt{Mor}(\mathcal{K}))$ of generic objects $A, B, \ldots$ in $\mathtt{Ob}(\mathcal{K})$ and generic arrows $f : A \to B$, $g : B \to C, \ldots$ in $\mathtt{Mor}(\mathcal{K})$ between objects, with *associative composition*:

$$A \xrightarrow{f} B \xrightarrow{g} C = A \xrightarrow{f \circ g} C,$$

and *identity* (*loop*) arrow. (Note that in topological literature, $\mathtt{Hom}(\mathcal{K})$ or $\mathtt{hom}(\mathcal{K})$ is used instead of $\mathtt{Mor}(\mathcal{K})$; see [Swi75]).

A category $\mathcal{K}$ is usually depicted as a *commutative diagram* (i.e., a diagram with a common *initial object* A and *final object* D):



To make this more precise, we say that a *category* $\mathcal{K}$ is defined if we have:

1. A *class of objects* $\{A, B, C, ...\}$ of $\mathcal{K}$, denoted by $\mathtt{Ob}(\mathcal{K})$;
2. A *set of morphisms*, or *arrows* $\mathtt{Mor}_{\mathcal{K}}(A, B)$, with elements $f : A \to B$, defined for any *ordered pair* $(A, B) \in \mathcal{K}$, such that for two different pairs $(A, B) \neq (C, D)$ in $\mathcal{K}$, we have $\mathtt{Mor}_{\mathcal{K}}(A, B) \cap \mathtt{Mor}_{\mathcal{K}}(C, D) = \emptyset$;
3. For any *triplet* $(A, B, C) \in \mathcal{K}$ with $f : A \to B$ and $g : B \to C$, there is a *composition* of morphisms

$$\mathtt{Mor}_{\mathcal{K}}(B, C) \times \mathtt{Mor}_{\mathcal{K}}(A, B) \ni (g, f) \to g \circ f \in \mathtt{Mor}_{\mathcal{K}}(A, C),$$

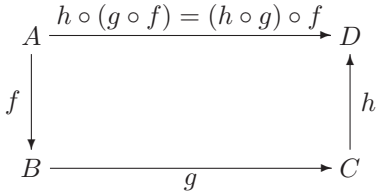
written schematically as

$$\frac{f : A \to B, \qquad g : B \to C}{g \circ f : A \to C}.$$

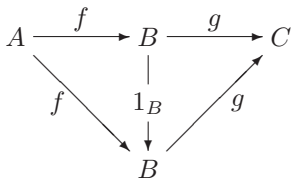
If we have a morphism $f \in \mathtt{Mor}_{\mathcal{K}}(A, B)$, (otherwise written $f : A \to B$, or $A \xrightarrow{\ f\ } B$), then $A = \mathrm{dom}(f)$ is a *domain* of f, and $B = \mathrm{cod}(f)$ is a *codomain* of f (of which *range* of f is a subset) and denoted $B = \mathrm{ran}(f)$.

To make $\mathcal{K}$ a category, it must also fulfill the following two properties:

1. *Associativity of morphisms*: for all $f \in \mathtt{Mor}_{\mathcal{K}}(A, B)$, $g \in \mathtt{Mor}_{\mathcal{K}}(B, C)$, and $h \in \mathtt{Mor}_{\mathcal{K}}(C, D)$, we have $h \circ (g \circ f) = (h \circ g) \circ f$; in other words, the following diagram is commutative



2. *Existence of identity morphism*: for every object $A \in \mathtt{Ob}(\mathcal{K})$ exists a unique identity morphism $1_A \in \mathtt{Mor}_{\mathcal{K}}(A, A)$; for any two morphisms $f \in \mathtt{Mor}_{\mathcal{K}}(A, B)$, and $g \in \mathtt{Mor}_{\mathcal{K}}(B, C)$, compositions with identity morphism $1_B \in \mathtt{Mor}_{\mathcal{K}}(B, B)$ give $1_B \circ f = f$ and $g \circ 1_B = g$, i.e., the following diagram is commutative:



The set of all morphisms of the category $\mathcal{K}$ is denoted

$$\mathtt{Mor}(\mathcal{K}) = \bigcup_{A,B \in Ob(\mathcal{K})} \mathtt{Mor}_{\mathcal{K}}(A,B).$$

If for two morphisms $f \in \mathtt{Mor}_{\mathcal{K}}(A,B)$ and $g \in \mathtt{Mor}_{\mathcal{K}}(B,A)$ the equality $g \circ f = 1_A$ is valid, then the morphism g is said to be *left inverse* (or *retraction*), of f, and f *right inverse* (or *section*) of g. A morphism which is both right and left inverse of f is said to be *two–sided inverse* of f.

A morphism $m : A \to B$ is called *monomorphism* in $\mathcal{K}$ (i.e., *one–to–one*, or *injection* map), if for any two parallel morphisms $f_1, f_2 : C \to A$ in $\mathcal{K}$ the equality $m \circ f_1 = m \circ f_2$ implies $f_1 = f_2$; in other words, m is monomorphism if it is *left cancellable*. Any morphism with a left inverse is monomorphism.

A morphism $e : A \to B$ is called *epimorphism* in $\mathcal{K}$ (i.e., *onto*, or *surjection* map), if for any two morphisms $g_1, g_2 : B \to C$ in $\mathcal{K}$ the equality $g_1 \circ e = g_2 \circ e$ implies $g_1 = g_2$; in other words, e is epimorphism if it is *right cancellable*. Any morphism with a right inverse is epimorphism.

A morphism $f : A \to B$ is called *isomorphism* in $\mathcal{K}$ (denoted as $f : A \cong B$) if there exists a morphism $f^{-1} : B \to A$ which is a two–sided inverse of f in $\mathcal{K}$. The relation of isomorphism is reflexive, symmetric, and transitive, i.e., equivalence relation.

For example, an isomorphism in the category of sets is called a set–isomorphism, or a *bijection*, in the category of topological spaces is called a topological isomorphism, or a *homeomorphism*, in the category of differentiable manifolds is called a differentiable isomorphism, or a *diffeomorphism*.

A morphism $f \in \mathtt{Mor}_{\mathcal{K}}(A,B)$ is *regular* if there exists a morphism $g : B \to A$ in $\mathcal{K}$ such that $f \circ g \circ f = f$. Any morphism with either a left or a right inverse is regular.

An object T is a *terminal object* in $\mathcal{K}$ if to each object $A \in \mathtt{Ob}(\mathcal{K})$ there is exactly one arrow $A \to T$. An object S is an *initial object* in $\mathcal{K}$ if to each object $A \in \mathtt{Ob}(\mathcal{K})$ there is exactly one arrow $S \to A$. A *null object* $Z \in \mathtt{Ob}(\mathcal{K})$ is an object which is both initial and terminal; it is unique up to isomorphism. For any two objects $A, B \in \mathtt{Ob}(\mathcal{K})$ there is a unique morphism $A \to Z \to B$ (the composite through Z), called the *zero morphism* from A to B.

A notion of subcategory is analogous to the notion of subset. A subcategory $\mathcal{L}$ of a category $\mathcal{K}$ is said to be a *complete subcategory* iff for any objects $A, B \in \mathcal{L}$, every morphism $A \to B$ of $\mathcal{L}$ is in $\mathcal{K}$.

The standard categories that we will use in this book are:

- $\mathcal{S}$ – all sets as objects and all functions between them as morphisms;

- $\mathcal{PS}$ – all pointed sets as objects and all functions between them preserving base point as morphisms;
- $\mathcal{V}$ – all vector spaces as objects and all linear maps between them as morphisms;
- $\mathcal{B}$ – Banach spaces over $\mathbb{R}$ as objects and bounded linear maps between them as morphisms;
- $\mathcal{G}$ – all groups as objects, all homomorphisms between them as morphisms;
- $\mathcal{A}$ – Abelian groups as objects, homomorphisms between them as morphisms;
- $\mathcal{AL}$ – all algebras (over a given field $\mathbb{K}$) as objects, all their homomorphisms between them as morphisms;
- $\mathcal{T}$ – all topological spaces as objects, all continuous functions between them as morphisms;
- $\mathcal{PT}$ – pointed topological spaces as objects, continuous functions between them preserving base point as morphisms;
- $\mathcal{TG}$ – all topological groups as objects, their continuous homomorphisms as morphisms;
- $\mathcal{M}$ – all smooth manifolds as objects, all smooth maps between them as morphisms;
- $\mathcal{M}_n$ – nD manifolds as objects, their local diffeomorphisms as morphisms;
- $\mathcal{LG}$ – all Lie groups as objects, all smooth homomorphisms between them as morphisms;
- $\mathcal{LAL}$ – all Lie algebras (over a given field $\mathbb{K}$) as objects, all smooth homomorphisms between them as morphisms;
- $\mathcal{TB}$ – all tangent bundles as objects, all smooth tangent maps between them as morphisms;
- $\mathcal{T}^*\mathcal{B}$ – all cotangent bundles as objects, all smooth cotangent maps between them as morphisms;
- $\mathcal{VB}$ – all smooth vector bundles as objects, all smooth homomorphisms between them as morphisms;
- $\mathcal{FB}$ – all smooth fibre bundles as objects, all smooth homomorphisms between them as morphisms;

A *groupoid* is a category in which every morphism is invertible. A typical groupoid is the *fundamental groupoid* $\Pi_1(X)$ of a topological space X. An object of $\Pi_1(X)$ is a point $x \in X$, and a morphism $x \to x'$ of $\Pi_1(X)$ is a homotopy class of paths f from x to x'. The *composition* of paths $g : x' \to x''$ and $f : x \to x'$ is the path h which is 'f followed by g'. Composition applies also to homotopy classes, and makes $\Pi_1(X)$ a category and a groupoid (the inverse of any path is the same path traced in the opposite direction).

A *group* is a groupoid with one object, i.e., a *category with one object* in which *all morphisms are isomorphisms*. Therefore, if we try to generalize the concept of a group, keeping associativity as an essential property, we get the notion of a category.

A category is *discrete* if every morphism is an identity. A *monoid* is a category with one object. A *group* is a category with one object in which every morphism has a two–sided inverse under composition.

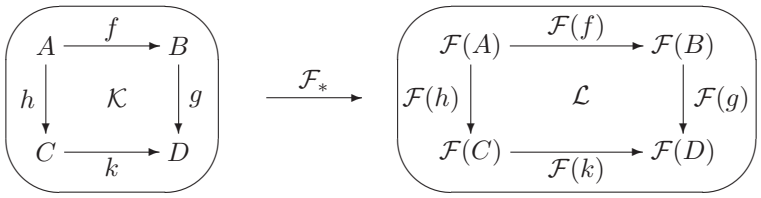
Homological algebra was the progenitor of category theory (see e.g., [Die88]). Generalizing L. Euler's formula $f + v = e + 2$ for the faces, vertices and edges of a convex polyhedron, E. Betti defined *numerical invariants of spaces* by formal addition and subtraction of faces of various dimensions; H. Poincaré formalized these and introduced homology. E. Noether stressed the fact that these calculations go on in Abelian groups, and that the operation ∂_n taking a face of dimension n to the alternating sum of faces of dimension $n-1$ which form its boundary is a homomorphism, and it also satisfies $\partial_n \cdot \partial_{n+1} = 0$. There are many ways of approximating a given space by polyhedra, but the quotient $H_n = \operatorname{Ker} \partial_n / \operatorname{Im} \partial_{n+1}$ is an invariant, the *homology group*. Since Noether, the groups have been the object of study instead of their dimensions, which are the *Betti numbers* (see Chapter 4 for details).

1.2.3 Functors

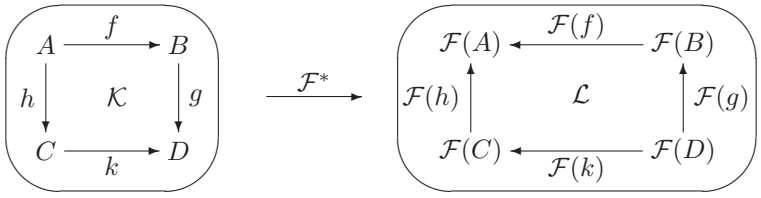
In algebraic topology, one attempts to assign to every topological space X some algebraic object $\mathcal{F}(X)$ in such a way that to every C^0-function $f : X \to Y$ there is assigned a homomorphism $\mathcal{F}(f) : \mathcal{F}(X) \longrightarrow \mathcal{F}(Y)$ (see [Swi75, DP97]). One advantage of this procedure is, e.g., that if one is trying to prove the non–existence of a C^0-function $f : X \to Y$ with certain properties, one may find it relatively easy to prove the non–existence of the corresponding algebraic function $\mathcal{F}(f)$ and hence deduce that f could not exist. In other words, $\mathcal{F}$ is to be a 'homomorphism' from one category (e.g., $\mathcal{T}$) to another (e.g., $\mathcal{G}$ or $\mathcal{A}$). Formalization of this notion is a *functor*.

A functor is a generic *picture* projecting one category into another. Let $\mathcal{K} = (\mathtt{Ob}(\mathcal{K}), \mathtt{Mor}(\mathcal{K}))$ be a *source* (or domain) *category* and $\mathcal{L} = (\mathtt{Ob}(\mathcal{L}), \mathtt{Mor}(\mathcal{L}))$ be a *target* (or codomain) category. A functor $\mathcal{F} = (\mathcal{F}_O, \mathcal{F}_M)$ is defined as a pair of maps, $\mathcal{F}_O : \mathtt{Ob}(\mathcal{K}) \to \mathtt{Ob}(\mathcal{L})$ and $\mathcal{F}_M : \mathtt{Mor}(\mathcal{K}) \to \mathtt{Mor}(\mathcal{L})$, preserving categorical symmetry (i.e., commutativity of all diagrams) of $\mathcal{K}$ in $\mathcal{L}$.

More precisely, a *covariant functor*, or simply a *functor*, $\mathcal{F}_* : \mathcal{K} \to \mathcal{L}$ is a *picture* in the target category $\mathcal{L}$ of (all objects and morphisms of) the source category $\mathcal{K}$:



Similarly, a *contravariant functor*, or a *cofunctor*, $\mathcal{F}^* : \mathcal{K} \to \mathcal{L}$ is a *dual picture* with reversed arrows:



In other words, a *functor* $\mathcal{F} : \mathcal{K} \to \mathcal{L}$ from a *source* category $\mathcal{K}$ to a *target* category $\mathcal{L}$, is a pair $\mathcal{F} = (\mathcal{F}_O, \mathcal{F}_M)$ of maps $\mathcal{F}_O : \mathtt{Ob}(\mathcal{K}) \to \mathtt{Ob}(\mathcal{L})$, $\mathcal{F}_M : \mathtt{Mor}(\mathcal{K}) \to \mathtt{Mor}(\mathcal{L})$, such that

1. If $f \in \mathtt{Mor}_{\mathcal{K}}(A, B)$ then $\mathcal{F}_M(f) \in \mathtt{Mor}_{\mathcal{L}}(\mathcal{F}_O(A), \mathcal{F}_O(B))$ in case of the *covariant* functor $\mathcal{F}_*$, and $\mathcal{F}_M(f) \in \mathtt{Mor}_{\mathcal{L}}(\mathcal{F}_O(B), \mathcal{F}_O(A))$ in case of the *contravariant* functor $\mathcal{F}^*$;
2. For all $A \in \mathtt{Ob}(\mathcal{K}) : \mathcal{F}_M(1_A) = 1_{\mathcal{F}_O(A)}$;
3. For all $f, g \in \mathtt{Mor}(\mathcal{K})$: if $\mathrm{cod}(f) = \mathrm{dom}(g)$, then $\mathcal{F}_M(g \circ f) = \mathcal{F}_M(g) \circ \mathcal{F}_M(f)$ in case of the *covariant* functor $\mathcal{F}_*$, and $\mathcal{F}_M(g \circ f) = \mathcal{F}_M(f) \circ \mathcal{F}_M(g)$ in case of the *contravariant* functor $\mathcal{F}^*$.

Category theory originated in algebraic topology, which tried to assign algebraic invariants to topological structures. The golden rule of such *invariants* is that they should be *functors*. For example, the *fundamental group* π_1 is a functor. Algebraic topology constructs a group called the *fundamental group* $\pi_1(X)$ from any topological space X, which keeps track of how many holes the space X has. But also, any map between topological spaces determines a homomorphism $\phi : \pi_1(X) \to \pi_1(Y)$ of the fundamental groups. So the fundamental group is really a functor $\pi_1 : \mathcal{T} \to \mathcal{G}$. This allows us to completely transpose any situation involving *spaces* and *continuous maps* between them to a parallel situation involving *groups* and *homomorphisms* between them, and thus reduce some topology problems to algebra problems.

Also, singular homology in a given dimension n assigns to each topological space X an Abelian group $H_n(X)$, its nth *homology group* of X, and also to each continuous map $f : X \to Y$ of spaces a corresponding homomorphism $H_n(f) : H_n(X) \to H_n(Y)$ of groups, and this in such a way that $H_n(X)$ becomes a functor $H_n : \mathcal{T} \to \mathcal{A}$.

The leading idea in the *use of functors in topology* is that H_n or π_n gives an algebraic picture or image not just of the topological spaces X, Y but also of all the continuous maps $f : X \to Y$ between them.

Similarly, there is a functor $\Pi_1 : \mathcal{T} \to \mathcal{G}$, called the 'fundamental groupoid functor', which plays a very basic role in algebraic topology. Here's how we get from any space X its 'fundamental groupoid' $\Pi_1(X)$. To say what the groupoid $\Pi_1(X)$ is, we need to say what its objects and morphisms are. The objects in $\Pi_1(X)$ are just the *points* of X and the morphisms are just certain equivalence classes of *paths* in X. More precisely, a morphism $f : x \to y$ in $\Pi_1(X)$ is just an equivalence class of continuous paths from x to y, where two

paths from x to y are decreed equivalent if one can be continuously deformed to the other while not moving the endpoints. (If this equivalence relation holds we say the two paths are 'homotopic', and we call the equivalence classes 'homotopy classes of paths' (see [MacL71, Swi75]).

Another examples are covariant *forgetful* functors:

- From the category of topological spaces to the category of sets; it 'forgets' the topology–structure.
- From the category of metric spaces to the category of topological spaces with the topology induced by the metrics; it 'forgets' the metric.

For each category $\mathcal{K}$, the *identity functor* $I_{\mathcal{K}}$ takes every $\mathcal{K}-$object and every $\mathcal{K}-$morphism to itself.

Given a category $\mathcal{K}$ and its subcategory $\mathcal{L}$, we have an *inclusion functor* $\mathtt{In} : \mathcal{K} \longrightarrow \mathcal{K}$.

Given a category $\mathcal{K}$, a *diagonal functor* $\Delta : \mathcal{K} \longrightarrow \mathcal{K}$ takes each object $A \in \mathcal{K}$ to the object (A, A) in the product category $\mathcal{K} \times \mathcal{K}$.

Given a category $\mathcal{K}$ and a category of sets $\mathcal{S}$, each object $A \in \mathcal{K}$ determines a *covariant Hom–functor* $\mathcal{K}[A, \_] : \mathcal{K} \to \mathcal{S}$, a *contravariant Hom–functor* $\mathcal{K}[\_, A] : \mathcal{K} \longrightarrow \mathcal{S}$, and a *Hom–bifunctor* $\mathcal{K}[\_, \_] : \mathcal{K}^{op} \times \mathcal{K} \to \mathcal{S}$.

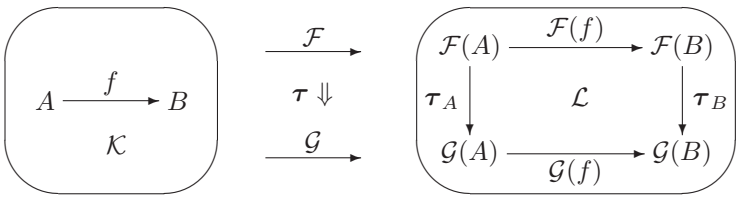
A functor $\mathcal{F} : \mathcal{K} \to \mathcal{L}$ is a *faithful functor* if for all $A, B \in \mathtt{Ob}(\mathcal{K})$ and for all $f, g \in \mathtt{Mor}_{\mathcal{K}}(A, B)$, $\mathcal{F}(f) = \mathcal{F}(g)$ implies $f = g$; it is a *full functor* if for every $h \in \mathtt{Mor}_{\mathcal{L}}(\mathcal{F}(A), \mathcal{F}(B))$, there is $g \in \mathtt{Mor}_{\mathcal{K}}(A, B)$ such that $h = \mathcal{F}(g)$; it is a *full embedding* if it is both full and faithful.

A *representation of a group* is a functor $\mathcal{F} : \mathcal{G} \to \mathcal{V}$.

Similarly, we can define a *representation of a category* to be a functor $\mathcal{F} : \mathcal{K} \to \mathcal{V}$ from the 2−category $\mathcal{K}$ (a 'big' category including all ordinary, or 'small' categories, see subsection (1.2.7) below) to the category of vector spaces $\mathcal{V}$. In this way, a category is a generalization of a group and group representations are a special case of category representations.

1.2.4 Natural Transformations

A *natural transformation* (i.e., a *functor morphism*) $\boldsymbol{\tau} : \mathcal{F} \stackrel{\cdot}{\to} \mathcal{G}$ is a *map between two functors of the same variance*, $(\mathcal{F}, \mathcal{G}) : \mathcal{K} \rightrightarrows \mathcal{L}$, preserving categorical symmetry:



More precisely, all functors of the same variance from a source category $\mathcal{K}$ to a target category $\mathcal{L}$ form themselves objects of the *functor category* $\mathcal{L}^{\mathcal{K}}$. Morphisms of $\mathcal{L}^{\mathcal{K}}$, called *natural transformations*, are defined as follows.

Let $\mathcal{F} : \mathcal{K} \to \mathcal{L}$ and $\mathcal{G} : \mathcal{K} \to \mathcal{L}$ be two functors of the same variance from a category $\mathcal{K}$ to a category $\mathcal{L}$. Natural transformation $\mathcal{F} \xrightarrow{\boldsymbol{\tau}} \mathcal{G}$ is a family of morphisms such that for all $f \in \mathtt{Mor}_{\mathcal{K}}(A, B)$ in the source category $\mathcal{K}$, we have $\mathcal{G}(f) \circ \boldsymbol{\tau}_A = \boldsymbol{\tau}_B \circ \mathcal{F}(f)$ in the target category $\mathcal{L}$. Then we say that the *component* $\boldsymbol{\tau}_A : \mathcal{F}(A) \to \mathcal{G}(A)$ *is natural in* A.

If we think of a functor $\mathcal{F}$ as giving a *picture* in the target category $\mathcal{L}$ of (all the objects and morphisms of) the source category $\mathcal{K}$, then a natural transformation $\boldsymbol{\tau}$ represents a set of morphisms mapping the picture $\mathcal{F}$ to another picture $\mathcal{G}$, preserving the commutativity of all diagrams.

An invertible natural transformation, such that all components $\boldsymbol{\tau}_A$ are isomorphisms) is called a *natural equivalence* (or, *natural isomorphism*). In this case, the inverses $(\boldsymbol{\tau}_A)^{-1}$ in $\mathcal{L}$ are the components of a natural isomorphism $(\boldsymbol{\tau})^{-1} : \mathcal{G} \xrightarrow{*} \mathcal{F}$. Natural equivalences are among the most important *metamathematical constructions* in algebraic topology (see [Swi75]).

For example, let $\mathcal{B}$ be the category of Banach spaces over $\mathbb{R}$ and bounded linear maps. Define $D : \mathcal{B} \to \mathcal{B}$ by taking $D(X) = X^* =$ Banach space of bounded linear functionals on a space X and $D(f) = f^*$ for $f : X \to Y$ a bounded linear map. Then D is a cofunctor. $D^2 = D \circ D$ is also a functor. We also have the identity functor $1 : \mathcal{B} \to \mathcal{B}$. Define $T : 1 \to D \circ D$ as follows: for every $X \in \mathcal{B}$ let $T(X) : X \to D^2 X = X^{**}$ be the *natural inclusion* – that is, for $x \in X$ we have $[T(X)(x)](f) = f(x)$ for every $f \in X^*$. T is a natural transformation. On the subcategory of finite–dimensional Banach spaces T is even a natural equivalence. The largest subcategory of $\mathcal{B}$ on which T is a natural equivalence is called the category of reflexive Banach spaces [Swi75].

As S. Eilenberg and S. MacLane first observed, 'category' has been defined in order to define 'functor' and 'functor' has been defined in order to define 'natural transformation' [MacL71]).

Compositions of Natural Transformations

Natural transformations can be *composed* in two different ways. First, we have an 'ordinary' composition: if $\mathcal{F}, \mathcal{G}$ and $\mathcal{H}$ are three functors from the source category $\mathcal{A}$ to the target category $\mathcal{B}$, and then $\boldsymbol{\alpha} : \mathcal{F} \dot{\to} \mathcal{G}$, $\boldsymbol{\beta} : \mathcal{G} \dot{\to} \mathcal{H}$ are two natural transformations, then the formula

$$(\beta \circ \alpha)_A = \beta_A \circ \alpha_A, \qquad \text{for all} \quad A \in \mathcal{A}, \tag{1.33}$$

defines a new natural transformation $\beta \circ \alpha : \mathcal{F} \dot{\to} \mathcal{H}$. This composition law is clearly associative and possesses a unit $1_{\mathcal{F}}$ at each functor $\mathcal{F}$, whose $\mathcal{A}$–component is $1_{\mathcal{F}\mathcal{A}}$.

Second, we have the *Godement product* of natural transformations, usually denoted by $*$. Let $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ be three categories, $\mathcal{F}, \mathcal{G}$, $\mathcal{H}$ and $\mathcal{K}$ be four

functors such that $(\mathcal{F},\mathcal{G}) : \mathcal{A} \rightrightarrows \mathcal{B}$ and $(\mathcal{H},\mathcal{K}) : \mathcal{B} \rightrightarrows \mathcal{C}$, and $\boldsymbol{\alpha} : \mathcal{F} \stackrel{\cdot}{\to} \mathcal{G}$, $\boldsymbol{\beta} : \mathcal{H} \stackrel{\cdot}{\to} \mathcal{K}$ be two natural transformations. Now, instead of (1.33), the Godement composition is given by

$$(\beta * \alpha)_A = \beta_{GA} \circ H(\alpha_A) = K(\alpha_A) \circ \beta_{FA}, \qquad \text{for all} \quad A \in \mathcal{A}, \tag{1.34}$$

which defines a new natural transformation $\beta * \alpha : \mathcal{H} \circ \mathcal{F} \stackrel{\cdot}{\to} \mathcal{K} \circ \mathcal{G}$.

Finally, the two compositions (1.33) and (1.33) of natural transformations can be combined as

$$(\delta * \gamma) \circ (\beta * \alpha) = (\delta \circ \beta) * (\gamma \circ \alpha),$$

where $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are three categories, $\mathcal{F},\mathcal{G}$, $\mathcal{H}$, $\mathcal{K}$, $\mathcal{L}$, $\mathcal{M}$ are six functors, and $\boldsymbol{\alpha} : \mathcal{F} \stackrel{\cdot}{\to} \mathcal{H}$, $\boldsymbol{\beta} : \mathcal{G} \stackrel{\cdot}{\to} \mathcal{K}$, $\boldsymbol{\gamma} : \mathcal{H} \stackrel{\cdot}{\to} \mathcal{L}$, $\boldsymbol{\delta} : \mathcal{K} \stackrel{\cdot}{\to} \mathcal{M}$ are four natural transformations.

Dinatural Transformations

Double natural transformations are called *dinatural transformations*. An *end of a functor* $S : C^{op} \times C \to X$ is a universal dinatural transformation from a constant e to S. In other words, an end of S is a pair $\langle e, \omega \rangle$, where e is an object of X and $\omega : e \stackrel{\cdot\cdot}{\to} S$ is a *wedge (dinatural) transformation* with the property that to every wedge $\beta : x \stackrel{\cdot\cdot}{\to} S$ there is a unique arrow $h : x \to e$ of B with $\beta_c = \omega_c h$ for all $a \in C$. We call ω the *ending wedge* with *components* ω_c, while the object e itself, by abuse of language, is called the end of S and written with integral notation as $\int_c S(c,c)$; thus

$$S(c,c) \stackrel{\omega_c}{\to} \int_c S(c,c) = e.$$

Note that the 'variable of integration' c appears twice under the integral sign (once contravariant, once covariant) and is 'bound' by the integral sign, in that the result no longer depends on c and so is unchanged if 'c' is replaced by any other letter standing for an object of the category C. These properties are like those of the letter x under the usual integral symbol $\int f(x)\,dx$ of calculus.

Every end is manifestly a limit – specifically, a limit of a suitable diagram in X made up of pieces like $S(b,b) \to S(b,c) \to S(c,c)$.

For each functor $T : C \to X$ there is an isomorphism

$$\int_c S(c,c) = \int_c Tc \cong \operatorname{Lim} T,$$

valid when either the end of the limit exists, carrying the ending wedge to the limiting cone; the indicated notation thus allows us to write any limit as an

integral (an end) without explicitly mentioning the dummy variable (the first variable c of S).

A functor $H : X \to Y$ is said to *preserve the end* of a functor $S : C^{op} \times C \to X$ when $\omega : e \ddot{\to} S$ an end of S in X implies that $H\omega : He \ddot{\to} HS$ is an and for HS; in symbols

$$H \int_c S(c,c) = \int_c HS(c,c).$$

Similarly, H *creates* the end of S when to each end $v : y \ddot{\to} HS$ in Y there is a unique wedge $\omega : e \ddot{\to} S$ with $H\omega = v$, and this wedge ω is an end of S.

The definition of the coend of a functor $S : C^{op} \times C \to X$ is dual to that of an end. A *coend* of S is a pair $\langle d, \zeta \rangle$, consisting of an object $d \in X$ and a wedge $\zeta : S \ddot{\to} d$. The object d (when it exists, unique up to isomorphism) will usually be written with an integral sign and with the bound variable c as superscript; thus

$$S(c,c) \overset{\zeta_c}{\to} \int^c S(c,c) = d.$$

The formal properties of coends are dual to those of ends. Both are much like those for integrals in calculus (see [MacL71], for technical details).

1.2.5 Limits and Colimits

In abstract algebra constructions are often defined by an abstract property which requires the existence of unique morphisms under certain conditions. These properties are called *universal properties*. The *limit* of a functor generalizes the notions of inverse limit and product used in various parts of mathematics. The dual notion, *colimit*, generalizes direct limits and direct sums. Limits and colimits are defined via universal properties and provide many examples of *adjoint functors*.

A *limit* of a covariant functor $\mathcal{F} : \mathcal{J} \to \mathcal{C}$ is an object L of $\mathcal{C}$, together with morphisms $\phi_X : L \to \mathcal{F}(X)$ for every object X of $\mathcal{J}$, such that for every morphism $f : X \to Y$ in $\mathcal{J}$, we have $\mathcal{F}(f)\phi_X = \phi_Y$, and such that the following *universal property* is satisfied: for any object N of $\mathcal{C}$ and any set of morphisms $\psi_X : N \to \mathcal{F}(X)$ such that for every morphism $f : X \to Y$ in $\mathcal{J}$, we have $\mathcal{F}(f)\psi_X = \psi_Y$, there exists precisely one morphism $u : N \to L$ such that $\phi_X u = \psi_X$ for all X. If $\mathcal{F}$ has a limit (which it need not), then the limit is defined up to a unique isomorphism, and is denoted by $\lim \mathcal{F}$.

Analogously, a *colimit* of the functor $\mathcal{F} : \mathcal{J} \to \mathcal{C}$ is an object L of $\mathcal{C}$, together with morphisms $\phi_X : \mathcal{F}(X) \to L$ for every object X of $\mathcal{J}$, such that for every morphism $f : X \to Y$ in $\mathcal{J}$, we have $\phi_Y \mathcal{F}(X) = \phi_X$, and such that the following universal property is satisfied: for any object N of $\mathcal{C}$ and any set of morphisms $\psi_X : \mathcal{F}(X) \to N$ such that for every morphism $f : X \to Y$ in $\mathcal{J}$, we have $\psi_Y \mathcal{F}(X) = \psi_X$, there exists precisely one morphism $u : L \to N$ such

that $u\phi_X = \psi_X$ for all X. The colimit of $\mathcal{F}$, unique up to unique isomorphism if it exists, is denoted by $\operatorname{colim}\mathcal{F}$.

Limits and colimits are related as follows: A functor $\mathcal{F} : \mathcal{J} \to \mathcal{C}$ has a colimit iff for every object N of $\mathcal{C}$, the functor $X \longmapsto Mor_{\mathcal{C}}(\mathcal{F}(X), N)$ (which is a covariant functor on the dual category $\mathcal{J}^{op}$) has a limit. If that is the case, then $Mor_{\mathcal{C}}(\operatorname{colim}\mathcal{F}, N) = \lim Mor_{\mathcal{C}}(\mathcal{F}(-), N)$ for every object N of $\mathcal{C}$.

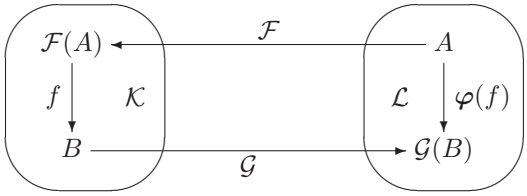
1.2.6 The Adjunction

The most important functorial operation is adjunction; as S. MacLane once said, "Adjoint functors arise everywhere" [MacL71].

The *adjunction* $\boldsymbol{\varphi} : \mathcal{F} \dashv \mathcal{G}$ between two functors $(\mathcal{F}, \mathcal{G}) : \mathcal{K} \leftrightarrows \mathcal{L}$ of *opposite variance* [Kan58], represents a *weak functorial inverse*

$$\frac{f : \mathcal{F}(A) \to B}{\boldsymbol{\varphi}(f) : A \to \mathcal{G}(B)}$$

forming a *natural equivalence* $\boldsymbol{\varphi} : \mathtt{Mor}_{\mathcal{K}}(\mathcal{F}(A), B) \xrightarrow{\boldsymbol{\varphi}} \mathtt{Mor}_{\mathcal{L}}(A, \mathcal{G}(B))$. The adjunction isomorphism is given by a *bijective correspondence* (a one–to–one and onto map on objects) $\boldsymbol{\varphi} : \mathtt{Mor}(\mathcal{K}) \ni f \to \boldsymbol{\varphi}(f) \in \mathtt{Mor}(\mathcal{L})$ of isomorphisms in the two categories, $\mathcal{K}$ (with a representative object A), and $\mathcal{L}$ (with a representative object B). It can be depicted as a (non–commutative) diagram



In this case $\mathcal{F}$ is called *left adjoint*, while $\mathcal{G}$ is called *right adjoint*.

In other words, an *adjunction* $F \dashv G$ *between two functors* $(\mathcal{F}, \mathcal{G})$ of opposite variance, from a source category $\mathcal{K}$ to a target category $\mathcal{L}$, is denoted by $(\mathcal{F}, \mathcal{G}, \boldsymbol{\eta}, \boldsymbol{\varepsilon}) : \mathcal{K} \leftrightarrows \mathcal{L}$. Here, $\mathcal{F} : \mathcal{L} \to \mathcal{K}$ is the *left (upper) adjoint functor*, $\mathcal{G} : \mathcal{L} \leftarrow \mathcal{K}$ is the *right (lower) adjoint functor*, $\boldsymbol{\eta} : 1_{\mathcal{L}} \to \mathcal{G} \circ \mathcal{F}$ is the *unit natural transformation* (or, *front adjunction*), and $\boldsymbol{\varepsilon} : \mathcal{F} \circ \mathcal{G} \to 1_{\mathcal{K}}$ is the *counit natural transformation* (or, *back adjunction*).

For example, $\mathcal{K} = \mathcal{S}$ is the category of sets and $\mathcal{L} = \mathcal{G}$ is the category of groups. Then $\mathcal{F}$ turns any set into the *free group* on that set, while the 'forgetful' functor $\mathcal{F}^*$ turns any group into the *underlying set* of that group. Similarly, all sorts of other 'free' and 'underlying' constructions are also left and right adjoints, respectively.

Right adjoints preserve *limits*, and left adjoints preserve *colimits*.

The category $\mathcal{C}$ is called a *cocomplete category* if every functor $\mathcal{F} : \mathcal{J} \to \mathcal{C}$ has a colimit. The following categories are cocomplete: $\mathcal{S}, \mathcal{G}, \mathcal{A}, \mathcal{T}$, and $\mathcal{PT}$.

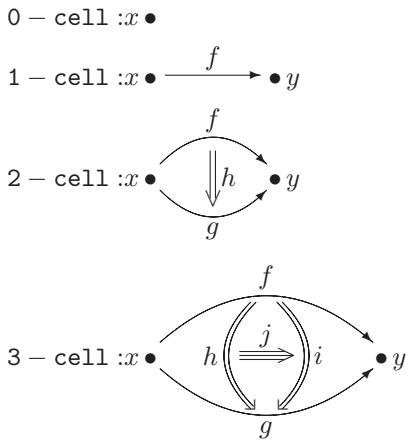
The importance of adjoint functors lies in the fact that every functor which has a left adjoint (and therefore is a right adjoint) is continuous. In the category $\mathcal{A}$ of Abelian groups, this e.g., shows that the kernel of a product of homomorphisms is naturally identified with the product of the kernels. Also, limit functors themselves are continuous. A covariant functor $\mathcal{F} : \mathcal{J} \to \mathcal{C}$ is *cocontinuous* if it transforms colimits into colimits. Every functor which has a right adjoint (and is a left adjoint) is cocontinuous.

The *analogy* between *adjoint functors* and *adjoint linear operators* relies upon a deeper analogy: just as in quantum theory the inner product $\langle \phi, \psi \rangle$ represents the *amplitude* to pass from ϕ to ψ, in category theory $Mor(A, B)$ represents the *set of ways* to go from A to B. These are to Hilbert spaces as categories are to sets. The analogues of adjoint linear operators between Hilbert spaces are certain adjoint functors between 2−Hilbert spaces [Bae97, BD98]. Similarly, the *adjoint representation* of a Lie group G is the linearized version of the action of G on itself by conjugation, i.e., for each $g \in G$, the inner automorphism $x \mapsto gxg^{-1}$ gives a linear transformation $Ad(g) : \mathfrak{g} \to \mathfrak{g}$, from the Lie algebra $\mathfrak{g}$ of G to itself.

1.2.7 $n-$Categories

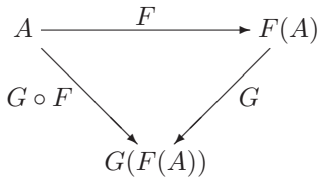
Generalization from 'Small' Categories to 'Big' $n-$Categories

If we think of a point in geometric space (either natural, or abstract) as an *object* (or, a 0−cell), and a path between two points as an *arrow* (or, a 1−*morphism*, or a 1−cell), we could think of a 'path of paths' as a 2−arrow (or, a 2−morphism, or a 2−cell), and a 'path of paths of paths' (or, a 3−morphism, or a 3−cell), etc. Here a 'path of paths' is just a continuous 1–parameter family of paths from between source and target points, which we can think of as tracing out a 2D surface, etc. In this way we get a 'skeleton' of an n−category, where a 1−category operates with 0−cells (objects) and 1−cells (arrows, causally connecting *source* objects with *target* ones), a 2−category operates with all the cells up to 2−cells [Ben67], a 3−category operates with all the cells up to 3−cells, etc. This skeleton clearly demonstrates the *hierarchical self–similarity* of *n–categories*:

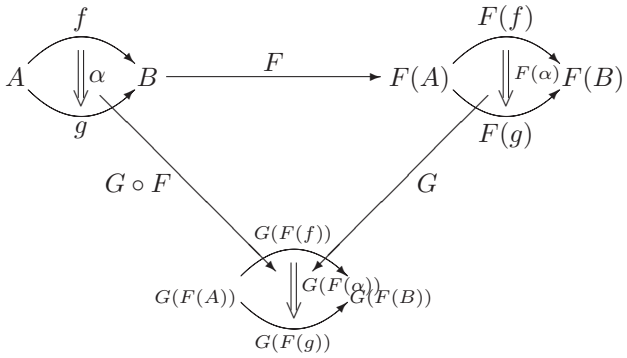


where triple arrow goes in the third direction, perpendicular to both single and double arrows. Categorical composition is defined by pasting arrows.

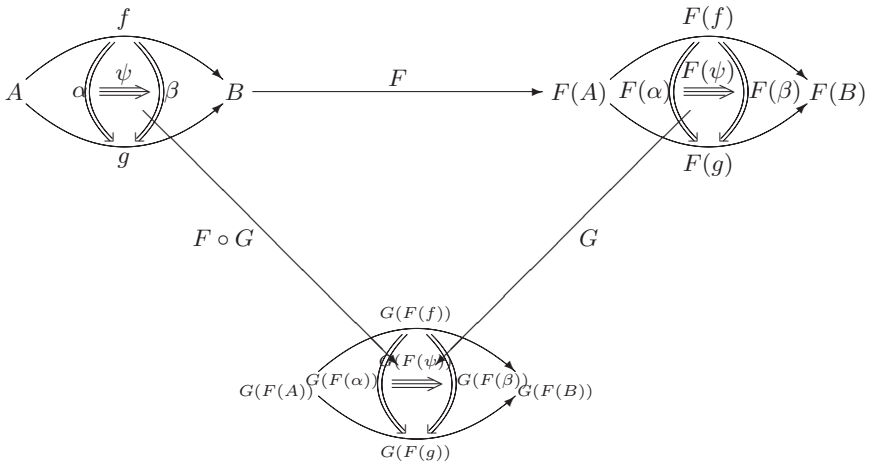
In this way defined, a 1−category can be depicted as a commutative triangle:



a 2−category is a commutative triangle:



a 3−category is a commutative triangle:



etc., up to $n-$categories.

Many deep–sounding results in mathematical sciences are obtained by the process of *categorification*[6] of the high school mathematics [CF94, BD98].

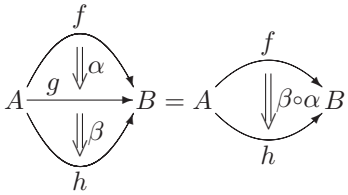
An $n-$category is a generic mathematical structure consisting of a collection of objects, a collection of arrows between objects, a collection of $2-$arrows between arrows [Ben67], a collection of $3-$arrows between $2-$arrows, and so on up to n [Bae97, BD98, Lei02, Lei03, Lei04].

More precisely, an $n-$category (for $n \geq 0$) consists of:

- $0-$cells, or objects, $A, B, \ldots$
- $1-$cells, or arrows, $A \xrightarrow{f} B$, with a composition

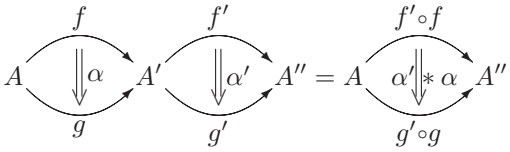
$$A \xrightarrow{f} B \xrightarrow{g} C = A \xrightarrow{g \circ f} C$$

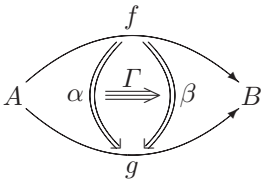
- $2-$cells, 'arrows between arrows', $A \overset{f}{\underset{g}{\Downarrow \alpha}} B$, with vertical compositions (denoted by $\circ$) and horizontal compositions (denoted by $*$), respectively given by



[6] Categorification means replacing sets with categories, functions with functors, and equations between functions by natural equivalences between functors. Iterating this process requires a theory of $n-$categories.

and

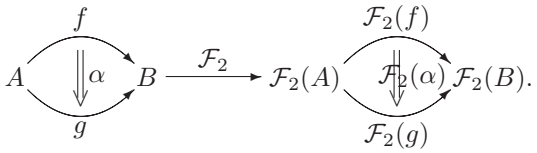


- 3−cells, 'arrows between arrows between arrows', 

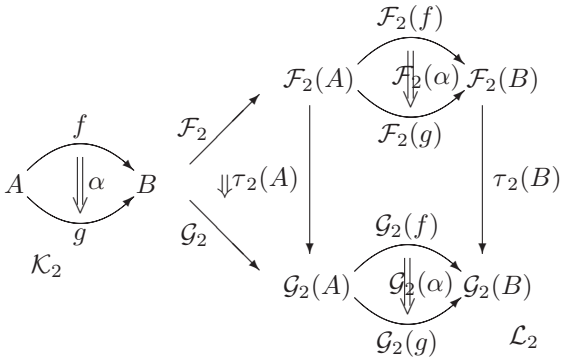
(where the $\Gamma-$arrow goes in a direction perpendicular to f and α), with various kinds of vertical, horizontal and mixed compositions,
- etc., up to $n-$cells.

Calculus of $n-$categories has been developed as follows. First, there is $\mathcal{K}_2$, the 2–category of all ordinary (or small) categories. $\mathcal{K}_2$ has categories $\mathcal{K}, \mathcal{L}, ...$ as objects, functors $\mathcal{F}, \mathcal{G} : \mathcal{K} \rightrightarrows \mathcal{L}$ as arrows, and natural transformations, like $\tau : \mathcal{F} \stackrel{\cdot}{\longrightarrow} \mathcal{G}$ as 2–arrows.

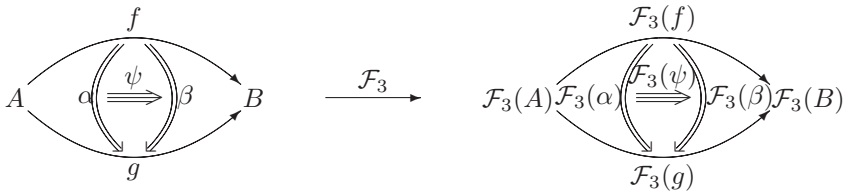
In a similar way, the arrows in a 3–category $\mathcal{K}_3$ are 2–functors $\mathcal{F}_2, \mathcal{G}_2, ...$ sending objects in $\mathcal{K}_2$ to objects in $\mathcal{L}_2$, arrows to arrows, and 2–arrows to 2–arrows, strictly preserving all the structure of $\mathcal{K}_2$



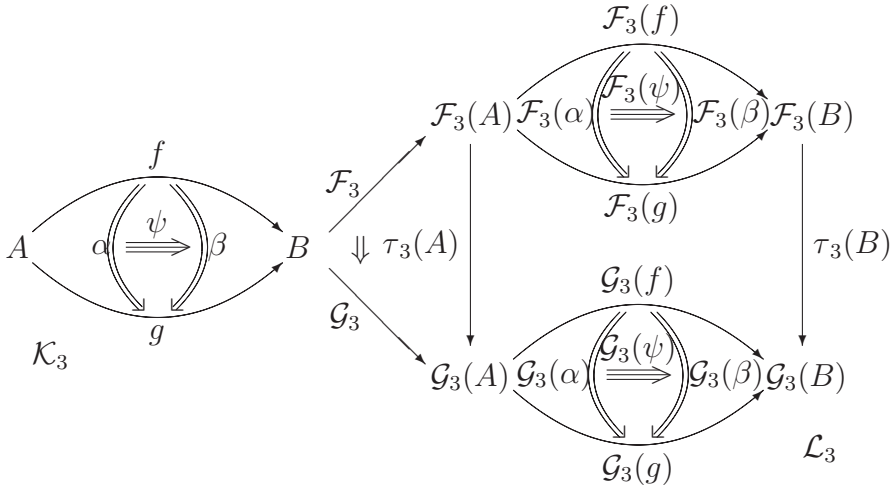
The 2–arrows in $\mathcal{K}_3$ are 2–natural transformations, like $\tau_2 : \mathcal{F}_2 \stackrel{2\cdot}{\Rightarrow} \mathcal{G}_2$ between 2–functors $\mathcal{F}_2, \mathcal{G}_2 : \mathcal{K}_2 \longrightarrow \mathcal{L}_2$ that sends each object in $\mathcal{K}_2$ to an arrow in $\mathcal{L}_2$ and each arrow in $\mathcal{K}_2$ to a 2–arrow in $\mathcal{L}_2$, and satisfies natural transformation–like conditions. We can visualize τ_2 as a prism going from one functorial picture of $\mathcal{K}_2$ in $\mathcal{L}_2$ to another, built using commutative squares:



Similarly, the arrows in a 4–category $\mathcal{K}_4$ are 3–functors $\mathcal{F}_3, \mathcal{G}_3, ...$ sending objects in $\mathcal{K}_3$ to objects in $\mathcal{L}_3$, arrows to arrows, and 2–arrows to 2–arrows, strictly preserving all the structure of $\mathcal{K}_3$



The 2–arrows in $\mathcal{K}_4$ are 3–natural transformations, like $\tau_3 : \mathcal{F} \overset{3.}{\Rightarrow} \mathcal{G}$ between 3–functors $\mathcal{F}_3, \mathcal{G}_3 : \mathcal{K}_3 \to \mathcal{L}_3$ that sends each object in $\mathcal{K}_3$ to a arrow in $\mathcal{L}_3$ and each arrow in $\mathcal{K}_3$ to a 2–arrow in $\mathcal{L}_3$, and satisfies natural transformation–like conditions. We can visualize τ_3 as a prism going from one picture of $\mathcal{K}_3$ in $\mathcal{L}_3$ to another, built using commutative squares:



Topological Structure of $n-$Categories

We already emphasized the topological nature of ordinary category theory. This fact is even more obvious in the general case of $n-$categories (see [Lei02, Lei03, Lei04]).

Homotopy Theory

Any topological manifold M gives rise to an $n-$*category* $\Pi_n(M)$ (its *fundamental* $n-$*groupoid*), in which 0–cells are *points* in M; 1–cells are *paths* in M (i.e., parameterized continuous maps $f : [0, 1] \to M$); 2–cells are *homotopies* (denoted by $\simeq$) *of paths* relative to endpoints (i.e., parameterized continuous

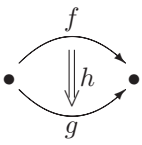
maps $h : [0,1] \times [0,1] \to M$); 3–cells are *homotopies of homotopies* of paths in M (i.e., parameterized continuous maps $j : [0,1] \times [0,1] \times [0,1] \to M$); categorical *composition* is defined by *pasting* paths and homotopies. In this way the following 'homotopy skeleton' emerges:

$\mathtt{0-cell} : x \bullet \qquad x \in M;$

$\mathtt{1-cell} : x \bullet \xrightarrow{f} \bullet y \qquad f : x \simeq y \in M,$

$f : [0,1] \to M,\ f : x \mapsto y,\ y = f(x),\ f(0) = x,\ f(1) = y;$

e.g., linear path: $f(t) = (1-t)x + ty;$

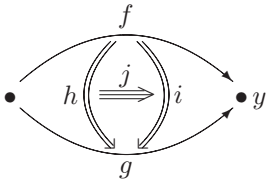
$\mathtt{2-cell} :$ 

$h : f \simeq g \in M,$

$h : [0,1] \times [0,1] \to M,\ h : f \mapsto g,\ g = h(f(x)),$

$h(x,0) = f(x),\ h(x,1) = g(x),\ h(0,t) = x,\ h(1,t) = y$

e.g., linear homotopy: $h(x,t) = (1-t)f(x) + tg(x);$

$\mathtt{3-cell} :$ 

$j : h \simeq i \in M,$

$j : [0,1] \times [0,1] \times [0,1] \to M,\ j : h \mapsto i,\ i = j(h(f(x)))$

$j(x,t,0) = h(f(x)),\ j(x,t,1) = i(f(x)),$

$j(x,0,s) = f(x),\ j(x,1,s) = g(x),$

$j(0,t,s) = x,\ j(1,t,s) = y$

e.g., linear composite homotopy: $j(x,t,s) = (1-t)\,h(f(x)) + t\,i(f(x)).$

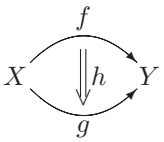
If M is a *smooth* manifold, then all included paths and homotopies need to be *smooth*. Recall that a *groupoid* is a category in which every morphism is invertible; its special case with only one object is a *group*.

Category $\mathcal{TT}$

Topological $n-$category $\mathcal{TT}$ has:

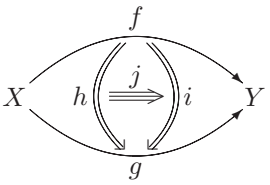
- 0–cells: topological spaces X
- 1–cells: continuous maps $X \xrightarrow{f} Y$

- 2–cells: homotopies h between f and $g : X \to Y$



  i.e., continuous maps $h : X \times [0,1] \to Y$, such that $\forall x \in X$, $h(x,0) = f(x)$ and $h(x,1) = g(x)$

- 3–cells: homotopies between homotopies : $X \to Y$

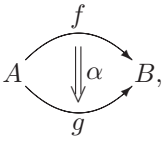


  i.e., continuous maps $j : X \times [0,1] \times [0,1] \to Y$.

Category $\mathcal{CK}$

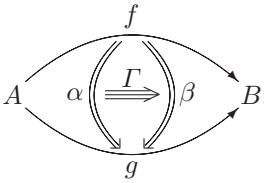
Consider an $n-$category $\mathcal{CK}$, which has:

- 0–cells: chain complexes A (of Abelian groups, say)
- 1–cells: chain maps $A \xrightarrow{f} B$
- 2–cells: chain homotopies $A \to B$,



  i.e., maps $\alpha : A \to B$ of degree 1

- 3–cells $A \to B$: homotopies between homotopies,



  i.e., maps $\Gamma : A \to B$ of degree 2 such that $d\Gamma - \Gamma d = \beta - \alpha$.

There ought to be some kind of map $\mathcal{CC} : \mathcal{TT} \Rightarrow \mathcal{CK}$ (see [Lei02, Lei03, Lei04]).

Categorification

Categorification is the process of finding category–theoretic analogs of set–theoretic concepts by replacing sets with categories, functions with functors, and equations between functions by natural isomorphisms between functors, which in turn should satisfy certain equations of their own, called 'coherence laws'. Iterating this process requires a theory of $n-$categories.

Categorification uses the following analogy between set theory and category theory [CF94, BD98]:

Set Theory	Category Theory
elements	objects
equations between elements	isomorphisms between objects
sets	categories
functions	functors
equations between functions	natural isomorphisms between functors

Just as sets have elements, categories have objects. Just as there are functions between sets, there are functors between categories. Now, the proper analog of an equation between elements is not an equation between objects, but an isomorphism. Similarly, the analog of an equation between functions is a natural isomorphism between functors.

1.2.8 Abelian Functorial Algebra

An *Abelian category* is a certain kind of category in which morphisms and objects can be added and in which kernels and cokernels exist and have the usual properties. The motivating prototype example of an Abelian category is the category of Abelian groups $\mathcal{A}$. Abelian categories are the framework for homological algebra (see [Die88]).

Given a homomorphism $f : A \to B$ between two objects $A \equiv \mathrm{Dom}\, f$ and $B \equiv \mathrm{Cod}\, f$ in an Abelian category $\mathcal{A}$, then its *kernel*, *image*, *cokernel* and *coimage* in $\mathcal{A}$ are defined respectively as:

$$\begin{aligned} \mathrm{Ker}\, f &= f^{-1}(e_B), & \mathrm{Coker}\, f &= \mathrm{Cod}\, f / \mathrm{Im}\, f, \\ \mathrm{Im}\, f &= f(A), & \mathrm{Coim}\, f &= \mathrm{Dom}\, f / \mathrm{Ker}\, f, \end{aligned}$$

where e_B is a unit of B [DP97].

In an Abelian category $\mathcal{A}$ a *composable* pair of arrows,

$$\bullet \xrightarrow{f} B \xrightarrow{g} \bullet$$

is *exact* at B iff $\mathrm{Im}\, f \equiv \mathrm{Ker}\, g$ (equivalence as subobjects of B) – or, equivalently, if $\mathrm{Coker}\, f \equiv \mathrm{Coim}\, g$ [MacL71].

For each arrow f in an Abelian category $\mathcal{A}$ the *triangular identities* read

$$\mathrm{Ker}(\mathrm{Coker}(\mathrm{Ker}\, f)) = \mathrm{Ker}\, f, \qquad \mathrm{Coker}(\mathrm{Ker}(\mathrm{Coker}\, f)) = \mathrm{Coker}\, f.$$

The diagram (with 0 the null object)

$$0 \longrightarrow A \xrightarrow{f} B \xrightarrow{g} C \longrightarrow 0 \tag{1.35}$$

is a *short exact sequence* when it is exact at A, at B, and at C.

Since $0 \to a$ is the zero arrow, exactness at A means just that f is *monic* (i.e., one–to–one, or injective map); dually, exactness at C means that g is *epic* (i.e., onto, or surjective map). Therefore, (1.35) is equivalent to

$$f = \mathrm{Ker}\, g, \qquad g = \mathrm{Coker}\, f.$$

Similarly, the statement that $h = \mathrm{Coker}\, f$ becomes the statement that the sequence

$$A \xrightarrow{f} B \xrightarrow{g} C \longrightarrow 0$$

is exact at B and at C. Classically, such a sequence was called a short right exact sequence. Similarly, $k = \mathrm{Ker}\, f$ is expressed by a short left exact sequence

$$0 \longrightarrow A \xrightarrow{f} B \xrightarrow{g} C.$$

If $\mathcal{A}$ and $\mathcal{A}'$ are Abelian categories, an *additive functor* $\mathcal{F} : \mathcal{A} \to \mathcal{A}'$ is a functor from $\mathcal{A}$ to $\mathcal{A}'$ with

$$\mathcal{F}(f + f') = \mathcal{F}f + \mathcal{F}f',$$

for any parallel pair of arrows $f, f' : b \to c$ in $\mathcal{A}$. It follows that $\mathcal{F}0 = 0$.

A functor $\mathcal{F} : \mathcal{A} \to \mathcal{A}'$ between Abelian categories $\mathcal{A}$ and $\mathcal{A}'$ is, by definition, *exact* when it preserves all finite limits and all finite colimits. In particular, an exact functor preserves kernels and cokernels, which means that

$$\mathrm{Ker}(\mathcal{F}f) = \mathcal{F}(\mathrm{Ker}\, f) \qquad \text{and} \qquad \mathrm{Coker}(\mathcal{F}f) = \mathcal{F}(\mathrm{Coker}\, f);$$

then $\mathcal{F}$ also preserves images, coimages, and carries exact sequences to exact sequences. By construction of limits from products and equalizers and dual constructions, $\mathcal{F} : \mathcal{A} \to \mathcal{A}'$ is exact iff it is additive and preserves kernels and cokernels.

A functor $\mathcal{F}$ is *left exact* when it preserves all finite limits. In other words, $\mathcal{F}$ is left exact iff it is additive and $\mathrm{Ker}(\mathcal{F}f) = \mathcal{F}(\mathrm{Ker}\, f)$ for all f: the last condition is equivalent to the requirement that $\mathcal{F}$ preserves short left exact sequences.

Similarly, a functor $\mathcal{F}$ is *right exact* when it preserves all finite colimits. In other words, $\mathcal{F}$ is right exact iff it is additive and $\mathrm{Coker}(\mathcal{F}f) = \mathcal{F}(\mathrm{Coker}\, f)$ for all f: the last condition is equivalent to the requirement that $\mathcal{F}$ preserves short right exact sequences.

In an Abelian category $\mathcal{A}$, a *chain complex* is a sequence

$$\ldots \longrightarrow c_{n+1} \xrightarrow{\partial_{n+1}} c_n \xrightarrow{\partial_n} c_{n-1} \longrightarrow \ldots$$

of composable arrows, with $\partial_n \partial_{n+1} = 0$ for all n. The sequence need not be exact at c_n; the deviation from exactness is measured by the nth *homology object*

$$H_n c = \mathrm{Ker}(\partial_n : c_n \longrightarrow c_{n-1}) / \mathrm{Im}(\partial_{n+1} : c_{n+1} \longrightarrow c_n).$$

Similarly, a *cochain complex* in an Abelian category $\mathcal{A}$ is a sequence

$$... \longrightarrow w_{n+1} \xrightarrow{d_{n+1}} w_n \xrightarrow{d_n} w_{n-1} \longrightarrow ...$$

of composable arrows, with $d_n d_{n+1} = 0$ for all n. The sequence need not be exact at w_n; the deviation from exactness is measured by the nth *cohomology object*

$$H^n w = \mathrm{Ker}(d_{n+1} : w_n \longrightarrow w_{n+1}) / \mathrm{Im}(d_n : w_{n-1} \longrightarrow w_n).$$

A *cycle* is a chain C such that $\partial C = 0$. A *boundary* is a chain C such that $C = \partial B$, for any other chain B.

A *cocycle* (a *closed form*) is a cochain ω such that $d\omega = 0$. A *coboundary* (an *exact form*) is a cochain ω such that $\omega = d\theta$, for any other cochain θ.