6

Covariant Biophysics of Electro–Muscular Stimulation

In this Chapter we develop covariant biophysics of electro–muscular stimulation, as an externally induced generator of our *covariant muscular forces*, $F_i = mg_{ij}a^i$. The so–called *functional electrical stimulation* (FES) of human skeletal muscles is used in rehabilitation and in medical orthotics to externally stimulate the muscles with damaged neural control (see, e.g., [VHI87]). However, the repetitive use of electro–muscular stimulation, besides functional, causes also structural changes in the stimulated muscles, giving the physiological effect of muscular training.

6.1 Basics of Electrical Muscular Stimulation

The use of low and very low frequency impulses in the body, delivered through electrodes, is known as transcutaneous stimulation of the nerves, electro–acupuncture and electro–stimulation. Here, an electromagnetic field accompanies the passage of the electric current through the conductive wire. This is generally known as the term 'electromagnetic therapy'.

In the original sense acupuncture meant the inserting of needles in specific regions of the body. Electro–acupuncture supplies the body with low–volt impulses through the medium of surface electrodes to specific body regions or by non specific electrodes. Transcutaneous electric stimulation of the nerves (TENS) has for years been a well known procedure in conventional medicine. The impulses that are produced with this type of stimulation, are almost identical with those of electro–stimulation, yet many doctors still assume, that they are two different therapies. This has resulted in TENS being considered as a daily therapy, while electro–acupuncture or electro–stimulation were treated as 'alternative therapy'. Apart from the fact, that electro–acupuncture electro–impulses are delivered through needles, both therapies should be considered identical. Patients, who have reservations about the use of needles, can by the use of electric impulses over surface electrodes on the skin, have a satisfactory alternative (see Figure 6.1). We choose the term electro–stimulation,

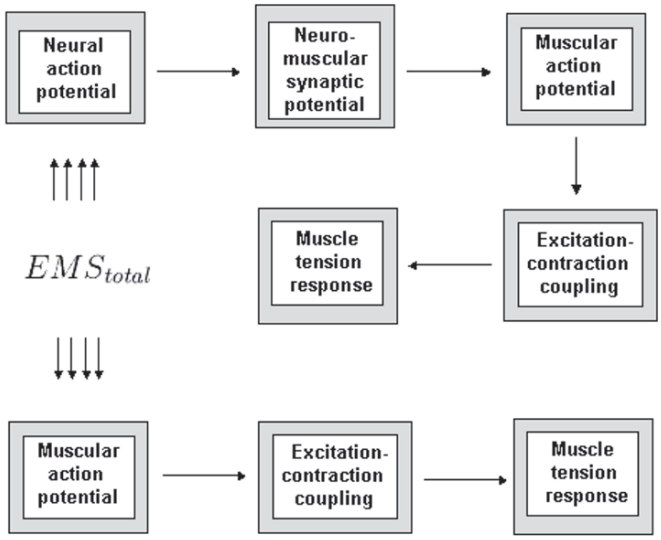

Fig. 6.1. Schematic of electrical muscular stimulation EMS.

because the acupuncture system is not included in all therapies. Clinical tests have showed that there are two specific types of reactions:

- The first reaction is spontaneous and dependent on the choice of body region. The stimulation of this part of the body results in an unloading, that can be compared with that of a battery. Normally this goes hand in hand with an immediate improvement in the patient. This effect of unloading may also be reached by non–specific electric stimulation.
- The second normal reaction is of a delayed nature, that results in relaxation and control of pain. Moreover two other important effects follow, that begin between 10 and 20 minutes after the start of the treatment. This reaction is associated (combined) with different chemicals, such as beta–endorphins and 5–hydrocytryptamins. When using low and very low frequency stimulations, the second effect is obtained by the utilization of specific frequencies on the body. This is independent of the choice of a specific part of the body, because the connected electromagnet makes the induction of secondary electric current in the whole body possible.

Now, when cosmetic surgeons perform *electrical muscular stimulation* (EMS, for short) on the human face or body (as schematically depicted on Figure 6.1), they usually take for granted half–a–dozen biophysical processes that are actually involved in this apparently simple *stimulus–response–type* action.

When the surface electrical muscular stimulation EMS pads are applied to the certain place of the human face or body, the first considerable tissue reaction is depolarization of sarcolemma, close to the electrodes. Muscular ability

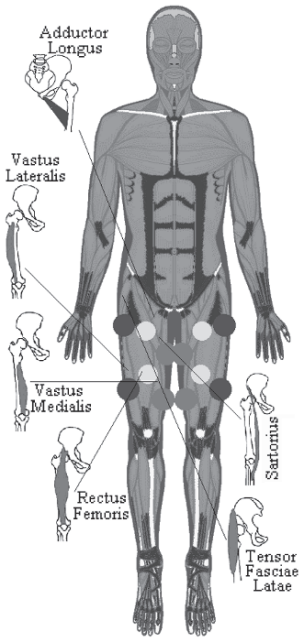

Fig. 6.2. Surface EMS performed on the human leg, together with anatomy of the stimulated leg muscles.

(similar to the neural one, but about ten times slower) to produce an action potential as a response to the stimulation, is termed 'excitability'. By means of the EMS, current is passed across a membrane to produce a transient depolarization of the resting potential of sarcolemma, which, if it is of sufficient duration and magnitude, can initiate the train of events that produces muscular action potential (see Figure 6.1). The minimum necessary intensity of stimulus is called the threshold stimulus. The term 'threshold' is commonly used to refer either to the absolute magnitude of the muscle–cell membrane potential at which an action potential is initiated or to the magnitude of depolarization from resting potential (in which the membrane naturally polarizes sodium and potassium ions) required to initiate an action potential. A stimulus of less than threshold intensity is refereed to as subthreshold, one of greater than threshold intensity as super–threshold. The threshold potential for excitation is not a fixed parameter. The thresholds of different muscle–membranes may vary considerably. Furthermore, the threshold of a single cell can change, either rapidly, as after a train of impulses produced by the EMS_{total}, or more slowly, in response to metabolic or hormonal influences.

On the other hand, in view of modern biophysics, there are *six distinctive phases of electrical muscular stimulation*, as usually performed using the surface pads (like those on Figures 6.1 and 6.1):

1. *Electrodynamic stimulation fields*, consisting of:
   a) External Maxwell electrodynamics (smooth, causal, unique and fully predictive); and
   b) Internal cellular bio–quantum electrodynamics (rapidly fluctuating, uncertain and stochastic, allowing only probabilistic approach).
2. *Muscular contraction paths*, consisting of:
   a) Anatomical external muscular mechanics; and
   b) Myofibrillar internal cellular bio–quantum mechanics.
3. *Geometric face & body shapes and curvatures*, consisting of:
   a) Smooth 2D external skin geometry; and
   b) Coarse–grained and fractal, internal nD cellular muscle–fat geometry.

Combined together, these six electro–mechano–geometric faces of electro–muscular stimulation generate the three–link $EMS-$*transition functor*:

$$\begin{matrix} ELECTRICAL \\ STIMUL \end{matrix} \Rightarrow \begin{matrix} MUSCULAR \\ CONTRACT \end{matrix} \Rightarrow \begin{matrix} FACE\,or\,BODY \\ SHAPE \end{matrix}$$

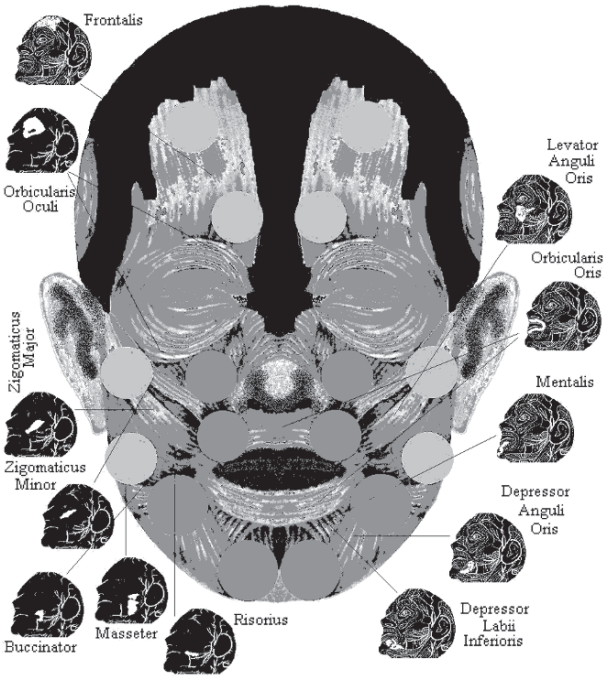

Fig. 6.3. Surface EMS performed on the human face, together with anatomy of the stimulated facial muscles (elaborated further in Table 6.5.1).

The EMS transition functor is based on Feynman–like experimental approach to electrical muscular stimulation (see Figure 6.2): the flow of elec-

tric current from the negative surface pad $A^{\ominus}$ to the positive pad $B^{\oplus}$ can be approximated by the vector sum of complex vectors $\sum_k \rho_k \, \mathrm{e}^{i\theta_k} = \rho_k(\cos\theta^k + i\sin\theta_k)$, (in the complex plane), where θ_k are proportional to the time taken by each vector ρ_k. This vector sum will be developed into the proper Feynman path integral (see [Fey98]).

The purpose of this Chapter is a modern and rigorous description of the above transition map, by elaboration of the six electro–mechano–geometric facets of the surface electro–muscular stimulation. All relevant classical biophysical theories are given in Appendix.

6.2 EMS Functor

Biophysically, electrical muscular stimulation represents a union of external electrical stimulation fields, internal myofibrillar excitation–contraction paths, and dissipative skin & fat geometries, formally written as

$$EMS_{total} = EMS_{fields} \bigcup EMS_{paths} \bigcup EMS_{geom}. \tag{6.1}$$

Following the current trends of the XXI century biophysics, corresponding to each of the three EMS–phases in (6.1) we formulate:

1. The *least action principle* (see section 3.3 above), to model a unique, external–anatomical, predictive and smooth, macroscopic EMS field–path–geometry; and
2. Associated *Feynman path integral* (see subsection 3.3.7 above), to model an ensemble of rapidly and stochastically fluctuating, internal, microscopic, fields–paths–geometries of the cellular EMS, to which the external–anatomical macro–level represents both time and ensemble *average*.[1]

In the proposed formalism, muscular excitation–contraction paths $x^i(t)$ are caused by electrodynamic stimulation fields $F^k(t)$, while they are both affected by dissipative and noisy skin & fat shapes and curvatures, defined by the local *Riemannian metric tensor* g_{ij}.

In the following text, we first formulate the global model for the EMS_{total}, to set up the general formalism to be specialized subsequently for each of the three EMS–phases.

6.2.1 Global macro–level of EMS_{total}

In general, at the *macroscopic* EMS–level we first formulate the *total action* $S[\Phi]$, our central quantity, which can be described through physical dimensions of $Energy \times Time = Effort$ (which is also the dimension of the *Planck constant* $\hbar$ $(= 1$ in normal units) (see, e.g., [DEF99]). This total action quantity

[1] Recall that *ergodic hypothesis* equates *time average* with *ensemble average*.

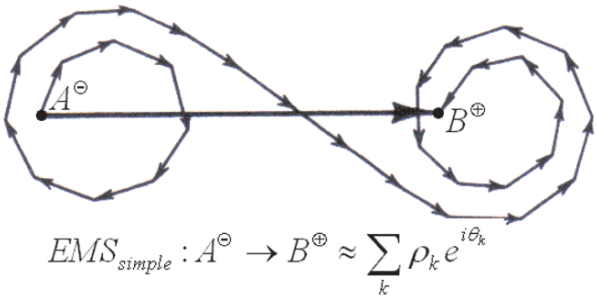

Fig. 6.4. Simplified Feynman–like experimental approach to electrical muscular stimulation: the flow of electric current from the negative surface pad $A^{\ominus}$ to the positive pad $B^{\oplus}$ can be approximated by the vector sum of complex vectors $\rho_k \, \mathrm{e}^{i\theta_k}$, where θ_k are proportional to the time taken by each vector ρ_k; this vector sum will be further developed into Feynman integral (see [Fey98]).

has immediate biophysical ramifications: *the greater the action – the higher the stimulation effect on the new shape.* The action $S[\Phi]$ depends on macroscopic fields, paths and geometries, commonly denoted by an abstract field symbol Φ^i. The action $S[\Phi]$ is formally defined as a temporal integral from the *initial* time instant t_{ini} to the *final* time instant t_{fin},

$$S[\Phi] = \int_{t_{ini}}^{t_{fin}} \mathfrak{L}[\Phi] \, dt, \tag{6.2}$$

with *Lagrangian density*, given by

$$\mathfrak{L}[\Phi] = \int d^n x \, \mathcal{L}(\Phi^i, \partial_{x^j} \Phi^i),$$

where the integral is taken over all n coordinates $x^j = x^j(t)$ of the EMS, and $\partial_{x^j} \Phi^i$ are time and space partial derivatives of the Φ^i-variables over coordinates.

Second, we formulate the *least action principle* as a minimal variation δ of the action $S[\Phi]$

$$\delta S[\Phi] = 0, \tag{6.3}$$

which, using variational *Euler–Lagrangian equations* (see section 3.3 above), derives field–motion–geometry of the unique and smooth $EMS-$transition functor

$$\mathcal{T} : STIMUL_{t_{ini}} \Rightarrow CONTRACT_{t_{mid}} \Rightarrow SHAPE_{t_{fin}},$$

acting at a macro–level from some initial time t_{ini} to the final time t_{fin} (via the intermediate time t_{mid}).

Here, we have in place *$n-$categorical Lagrangian–field structure* on the *muscular Riemannian configuration manifold* M,

$$\Phi^i : [0,1] \to M,\ \Phi^i : \Phi^i_0 \mapsto \Phi^i_1,$$

generalized from the recursive homotopy dynamics (3.5.1) above, using

$$\frac{d}{dt} f_{\dot{x}^i} = f_{x^i} \longrightarrow \partial_\mu \left(\frac{\partial \mathcal{L}}{\partial_\mu \Phi^i} \right) = \frac{\partial \mathcal{L}}{\partial \Phi^i},$$

with

$$[x_0, x_1] \rightarrowtail [\Phi^i_0, \Phi^i_1].$$

In this way, we get macro–objects in the global EMS: a single electrodynamic stimulation field described by Maxwell field equations, a single muscular excitation–contraction path described by Lagrangian equation of motion, and a single Riemannian skin & fat geometry.

6.2.2 Local Micro–Level of EMS_{total}

After having properly defined macro–level EMS_{total}, with a unique and globally–smooth $EMS-$transition functor $\mathcal{T}$, we move down to the *microscopic cellular* EMS–level of rapidly fluctuating electrodynamic fields, sarcomere–contraction paths and coarse–grained, fractal muscle–fat geometry, where we cannot define a unique and smooth field–path–geometry. The most we can do at this level of *fluctuating noisy uncertainty*, is to formulate an adaptive path integral and calculate overall probability amplitudes for ensembles of local transitions from negative EMS–pad $A^{\ominus}$ to the positive pad $B^{\oplus}$ (see Figure 6.2). This *probabilistic transition micro–dynamics* is given by a multi field–path–geometry, defining the microscopic *transition amplitude* corresponding to the macroscopic $EMS-$transition functor $\mathcal{T}$. So, what is externally the transition functor, internally is the transition amplitude. The absolute square of the transition amplitude is the *transition probability*.

Now, the total $EMS-$transition amplitude, from the initial state $STIMUL$, to the final state $SHAPE$, is defined on EMS_{total}[2]

$$\langle SHAPE | STIMUL \rangle_{total} \equiv \oint\!\!\!\!\!\!- : STIMUL_{t_0} \Rrightarrow SHAPE_{t_1}, \qquad (6.4)$$

given by *modern adaptive generalization* of the classical *Feynman path integral*, see [FH65, Fey72, Fey98, DEF99]). The transition map (6.4) calculates *overall probability amplitude* along a multitude of wildly fluctuating fields, paths and geometries, performing the *microscopic* transition from the micro–state $STIMUL_{t_0}$ occurring at initial micro–time instant t_0 to the micro–state $SHAPE_{t_1}$ at some later micro–time instant t_1, such that all micro–time instants fit inside the global transition interval $t_0, t_1, ..., t_s \in [t_{ini}, t_{fin}]$. It is symbolically written as

[2] We use the famous Dirac symbol $\oint\!\!\!\!\!\!-$ to denote summation over 'discrete spectrum' and integration over 'continuous spectrum' of fields, paths and geometries in the micro–EMS.

$$\langle SHAPE|STIMUL\rangle_{total} = \oint \mathcal{D}[w\Phi]\, \mathrm{e}^{\mathrm{i}S[\Phi]}, \tag{6.5}$$

where the Lebesgue integration is performed over all continuous $\Phi^i_{con} = fields + paths + geometries$, while summation is performed over all discrete processes and regional topologies Φ^j_{dis}. The symbolic differential $\mathcal{D}[w\Phi]$ in the general path integral (6.5), represents an *adaptive path measure*, defined as a weighted product

$$\mathcal{D}[w\Phi] = \lim_{N\longrightarrow\infty} \prod_{s=1}^{N} w_s d\Phi^i_s, \qquad (i = 1, ..., n = con + dis), \tag{6.6}$$

which is in practice satisfied with a large N.

In the exponent of the path integral (6.5) we have the action $S[\Phi]$ and the imaginary unit $\mathrm{i} = \sqrt{-1}$ (i can be converted into the real number -1 using the so–called *Wick rotation*). Feynman path integrals are usually computed by the use of *perturbative expansion methods* (see Appendix, section A.3; for other non–standard applications of Feynman path integrals see [Ing97, Ing98]).

In this way, we get a range of micro–objects in the local EMS_{total} at the short time–level: ensembles of rapidly fluctuating, noisy and crossing electrical stimulation fields, myofibrillar contraction paths and local skin & fat shape–geometries. However, by averaging process, both in time and along ensembles of fields, paths and geometries, we can recover the corresponding global, smooth and fully predictive, external EMS_{total} transition–dynamics $\mathcal{T}$.

6.2.3 Micro–Level Adaptation and Muscular Training

The adaptive path integral (6.5–6.6) incorporates the local muscular training process (see Appendix) according to the basic learning formula (see e.g., [Gro82, IJB99a])

$$NEW\ VALUE = OLD\ VALUE + INNOVATION,$$

where the term $VALUE$ represents respectively *biological images* of the $STIMUL$, $CONTRACT$ and $SHAPE$.

The general *synaptic weights* $w_s = w_s(t)$ in (6.6) are updated by the *homeostatic neuro–muscular feedbacks* during the transition process $\mathcal{T}$, according to one of the two standard neural training schemes, in which the micro–time level is traversed in discrete steps, i.e., if $t = t_0, t_1, ..., t_s$ then $t + 1 = t_1, t_2, ..., t_{s+1}$:

1. A *self–organized*, *unsupervised*, e.g., Hebbian–like training rule [Heb49]:

$$w_s(t + 1) = w_s(t) + \frac{\sigma}{\eta}(w^d_s(t) - w^a_s(t)), \tag{6.7}$$

   where $\sigma = \sigma(t)$, $\eta = \eta(t)$ denote *signal* and *noise*, respectively, while superscripts d and a denote *desired* and *achieved* muscular micro–states, respectively; or

2. A certain form of a *supervised gradient descent training*:

$$w_s(t+1) \;=\; w_s(t) - \eta \nabla J(t), \tag{6.8}$$

where η is a small constant, called the *step size*, or the *training rate,* and $\nabla J(n)$ denotes the gradient of the 'performance hyper–surface' at the $t-$th iteration.

6.3 Electrical Stimulation Fields: EMS_{fields}

6.3.1 External Smooth Maxwell Electrodynamics

On the macro–level in the phase EMS_{fields} we formulate the *electrodynamic field action principle* (see, e.g. [DEF99])

$$\delta S[F] = 0, \tag{6.9}$$

with the action $S[F]$ dependent on N electrodynamic stimulation fields $F^i = F^i(x)$, defined as a temporal integral

$$S[F] = \int_{t_{ini}}^{t_{fin}} \mathfrak{L}[F]\, dt, \tag{6.10}$$

with Lagrangian density given by

$$\mathfrak{L}[F] = \int d^n x\, \mathcal{L}(F_i, \partial_{x^j} F^i),$$

where the integral is taken over all n coordinates $x^j = x^j(t)$ of the EMS, and $\partial_{x^j} F^i$ are partial derivatives of the electrodynamic field variables over coordinates.

The action principle (6.9) implies the following *Maxwell electrodynamics*, presented here in vector, tensor and modern exterior differential form.

Given the following 3D vector–fields: the *electrical field* $\mathbf{E}$, the *magnetic field* $\mathbf{B}$ and the *electrical current* $\mathbf{J}$, as well as the *scalar electrical potential* ρ, the Maxwell electrical vector equations read[3] (see, e.g., [MTW73]):

1. *Electrostatics*:

$$\nabla \cdot \mathbf{E} \equiv \operatorname{div} \mathbf{E} = 4\pi\rho, \qquad \text{and}$$

2. *Electrodynamics*:

$$\partial_t \mathbf{E} - \nabla \times \mathbf{B} \equiv \partial_t \mathbf{E} - \operatorname{curl} \mathbf{B} = -4\pi \mathbf{J}.$$

[3] Only electrodynamic half of the Maxwell electro–magnetic field is elaborated here, as the other, magnetodynamic part has a minor role in physiology of electro–muscular stimulation.

Otherwise, given the 4D electromagnetic tensor–field *Faraday*,

$$F^{\alpha\beta} = \begin{pmatrix} 0 & E_x & E_y & E_z \\ -E_x & 0 & B_z & -B_y \\ -E_y & -B_z & 0 & B_x \\ -E_z & B_y & -B_x & 0 \end{pmatrix},$$

together with the 4D *electric current* vector–field $J^\alpha = (\mathbf{J}, -\rho)$, the tensor Maxwell equation reads (with *electrostatics and electrodynamics combined*):

$$F^{\alpha\beta}{}_{;\beta} = 4\pi J^\alpha.$$

Finally, given the two–form *Maxwell* $*F \equiv F_{\alpha\beta}$, which is a dual of the *Faraday* tensor, (also calculated as $F = dA$, where A is the one–form of *electrical potential*), and the three–form *charge* $*J$ (which is the dual one–form), the *exterior Maxwell equation* reads:

$$d*F = 4\pi*J.$$

The two–form *Maxwell* $*F \equiv F_{\alpha\beta}$, defines the *Lorentz force* one–form of the electro–muscular stimulation field,

$$Q_\alpha \equiv \dot{p}_\alpha \propto eF_{\alpha\beta}v^\beta,$$

where e is total electric charge and v^β is the velocity vector–field of the stimulation flow. This equation says that the muscular force Q_α generated by the simulation is proportional to the stimulation field strength $F_{\alpha\beta}$, velocity of the stimulation flow v^β through the skin–fat–muscle tissue, as well as the total stimulation charge e.

Now, let M be a smooth nD closed manifold with a Riemannian metric g_{ij} (see Chapter 2) and also with an exact two–form $F = dA$. Consider the problem of existence of closed extremals of the functional

$$S(\gamma) = \int_\gamma (\sqrt{g_{ij}\dot{x}^i\dot{x}^j} + A_i\dot{x}^i)\, dt,$$

on the space of closed curves $\gamma \in M$. This functional is a natural generalization of the usual functional of length, and its closed extremals correspond to periodic trajectories of the motion of particles on the Riemannian manifold M when the kinetic energy is defined by the metric tensor g_{ij}. When the Lagrangian function

$$L = \sqrt{g_{ij}\dot{x}^i\dot{x}^j} + A_i\dot{x}^i$$

is everywhere positive, we obtain the *Finsler metric* [Bah88], and the periodic problem can be studied by the methods of Morse theory (section 4.2.1 above).

6.3.2 Internal Cellular Bio–Quantum Electrodynamics

At the same time, on the micro–level in the phase EMS_{fields} we have the Feynman–type *sum over fields* F^i $(i = 1, ..., N)$ given by the adaptive path integral (see, e.g. [DEF99])

$$STIMUL_{fields} = \int \mathcal{D}[wF]\, \mathrm{e}^{\mathrm{i}S[F]} \quad \underrightarrow{Wick} \quad \int \mathcal{D}[wF]\, \mathrm{e}^{-S[F]}, \tag{6.11}$$

with action $S[F]$ given by the temporal integral (6.10), while $\underrightarrow{Wick}$ denotes the so–called Wick–rotation of the time variable t to imaginary values $t \mapsto \tau = \mathrm{i}t$. The resulting bio–quantum field represents the *bundle of cellular electrodynamic flux tubes*.

Now, during the XX century, the *electrodynamic flux tubes* were described by the Dirac–Schwinger–Tomonaga equations of *quantum electrodynamics*[4]. Today, the similar kind of *flux tubes* is in a more sophisticated way described by the *conformal Landau–Ginzburg model* (see, e.g., [DEF99]).

Technical details of these advanced physical theories are beyond the scope of the present article. In simplified terms, we can say that they all describe *field–generated solitons*[5] (see the next section for a *solitary model of muscular excitation–contraction*). The main point of all these quantum field theories and their biophysical applications is that their *macro–level averaging lift* (either in time or across the ensemble of cellular tubes) produces the classical Maxwell electrodynamics (6.3.1) above. On their own, they describe *rapidly fluctuating, fractal and noisy, electrodynamic fields flowing from the* ***source*** $(-)$ *electrode to the* ***sink*** $(+)$ *electrode* – as described in the *Schwinger formalism* (see, e.g., his lecture in the Nobel e–Museum).

[4] Mathematically, quantum electrodynamics has the structure of an Abelian *gauge theory* with a $U(1)$ gauge group. The gauge field which mediates the interaction between the charged spin 1/2 fields is the electromagnetic field. Physically, this translates to the picture of charged particles interacting with each other by the exchange of photons.

[5] In classical mathematical physics, by a *soliton* one usually means a "travelling wave" solution of a nonlinear partial differential equation $u_t = F(u, u_x, ...)$, i.e., a solution of the form $u(x,t) = f(x - vt)$. Solitons play a very important role in the theory of *integrable mechanical systems*, where any solution can be approximated by a superposition of solitons moving at different velocities, as we have in the next section where we develop a *solitary model of muscular excitation–contraction*. As a result, the theory of integrable systems is sometimes called soliton theory. In this section, however, we are interested in solitons arising in electrical field theory (as travelling wave solutions of the classical field equations) and primarily in the role they play in quantization of electrical field theories, which is a different point of view from the one in classical soliton theory.

6.4 Stimulated Muscular Contraction Paths: EMS_{paths}

6.4.1 External Anatomical Muscular Mechanics

On the macro–level in the phase EMS_{paths} we have the *muscular contraction action principle*

$$\delta S[x] = 0,$$

with the *mechanical action* $S[x]$ given by

$$S[x] = \int_{t_{ini}}^{t_{fin}} dt\,[\frac{m}{2}\dot{x}_i^2 + F(x)], \tag{6.12}$$

where overdot denotes time derivative, so that $\dot{x}^i$ represents the external (anatomical) muscular contraction speed, while m denotes the total estimated mass of the stimulated muscle. The corresponding Euler–Lagrangian equation, with the *kinetic energy of muscular contraction*

$$E_{kin} = \frac{1}{2} g_{ij}\,\dot{x}^i \dot{x}^j,$$

generated by *muscular Riemannian metrics* g_{ij} (see the next section on anatomical geometry), gives the *Newtonian equation of motion* (see e.g., [Arn89])

$$\frac{d}{dt}\frac{\partial E_{kin}}{\partial \dot{x}^i} - \frac{\partial E_{kin}}{\partial x^i} \equiv m\ddot{x}^i = -\partial_x F(x), \tag{6.13}$$

where ∂_s denotes the partial derivative with respect to the variable s (which is either space coordinate x or time t).

6.4.2 Internal Myofibrillar Bio–Quantum Mechanics

At the same time, on the micro–level in the phase EMS_{paths}, instead of a single path defined by the Newtonian equation of motion for the whole muscle (6.13), we have an ensemble of fluctuating and crossing, fractal paths with weighted probabilities (of the unit total sum). This ensemble of micro–paths is defined by the simplest instance of our adaptive path integral (6.5), similar to the Feynman's original *sum over histories*,

$$CONTRACT_{paths} = \int \mathcal{D}[wx]\,\mathrm{e}^{\mathrm{i}S[x]}, \tag{6.14}$$

where $\mathcal{D}[wx]$ is a functional measure on the *space of all weighted paths*, and the exponential depends on the action $S[x]$ given by (6.12). In the language of transition–propagators, the integral over histories (6.14) can be decomposed into the product of *myofibrillar action propagators*.[6] This procedure can be

[6] *Feynman propagators* are otherwise called *Fredholm kernels* or *Green's functions*.

redefined in a mathematically cleaner way if we Wick–rotate the time variable t to imaginary values $t \mapsto \tau = \mathrm{i}t$, thereby making all integrals real:

$$\int \mathcal{D}[wx]\, \mathrm{e}^{\mathrm{i}S[x]} \quad \underrightarrow{Wick} \quad \int \mathcal{D}[wx]\, \mathrm{e}^{-S[x]}. \tag{6.15}$$

Discretization of (6.15) gives the standard *thermodynamic partition function*

$$Z = \sum_j \mathrm{e}^{-w_j E^j/T}, \tag{6.16}$$

where E^j is the motion energy eigenvalue (reflecting each possible motivational energetic state), T is the temperature environmental control parameter, and the sum runs over all motion energy eigenstates (labelled by the index j). From (6.16), we can further calculate all thermodynamical and statistical EMS–properties (see [Fey72]), as for example, *transition entropy* $S = k_B \ln Z$, etc.

Now, both the action integral (6.12) and the path integral (6.14) are closely related to the *molecular soliton model of muscular contraction*, as described by the *Korteveg–De Vries equation* (3.32) and *nonlinear Schrödinger equation* (3.30) (see subsection 3.2.3 above). It is clear that these two solitary equations have a *quantum–mechanical* origin.[7] Recall, that by the use of the first quantization method (instead of the Feynman integral), every classical biodynamic observable F is represented in the Hilbert space $L^2(\psi)$ of square–integrable complex $\psi-$functions by a Hermitian (self–adjoint) linear operator $\hat{F}$ with real eigenvalues. The classical Poisson bracket $\{F, G\} = K$ corresponds to the *Dirac quantum commutator* $[\hat{F}, \hat{G}] = i\hat{K}$ (where, as always we have used normal units in which $\hbar = 1$). Therefore the classical evolution equation (3.28) corresponds, in the *Heisenberg picture*, to the *quantum evolution equation* (see, e.g., [Dir30])

$$i\hat{\dot{F}} = [\hat{F}, \hat{H}],$$

for any representative operator $\hat{F}$ and *quantum Hamiltonian operator* $\hat{H}$. By Ehrenfest's theorem (see, e.g., [Fey72]), this equation is also valid for expectation values $< \cdot >$ of observables, that is,

$$i < \hat{\dot{F}} > = < [\hat{F}, \hat{H}] > .$$

For technical details on *classical muscular mechanics*, including the celebrated work of Nobel Laureates:

1. the *microscopic sliding filament model* of A.F. Huxley;

[7] As Richard Feynman says in The Feynman Lectures on Physics: "Where did we get that [Schrödinger equation] from? It's not possible to derive it from anything you know. It came out of the mind of Schrödinger." Yet, Schrödinger equation can be (and usually is) derived from the Feynman path integral.

2. the *macroscopic force–velocity model* of A.V. Hill;
3. the celebrated *Hodgkin–Huxley neural* (and subsequently muscular) *excitation model* (A.L. Hodgkin and A.F. Huxley); and
4. the *Eccles model of synaptic activation*;

– see Appendix.

6.5 Anatomical Geometry of the Face & Body Shape: EMS_{geom}

6.5.1 External Face & Body Geometry

On the macro–level in the phase EMS_{geom} representing a smooth skin 2D manifold–patch M_2 with the Riemannian metric tensor $g_{ij} = g_{ij}(M_2)$ defined at each local face or body point, we formulate the *geometric action principle*

$$\delta S[g_{ij}] = 0,$$

where $S = S[g_{ij}]$ is the 2D *geodesic action* on the surface M,

$$S[g_{ij}] = \int d^n x \sqrt{g_{ij}\, dx^i dx^j}, \tag{6.17}$$

(Einstein's summation convention over repeated indices is assumed).
The corresponding Euler–Lagrangian equation gives the *geodesic equation* of the *shortest path* on the manifold M_2,

$$\ddot{x}^i + \Gamma^i_{jk}\, \dot{x}^j\, \dot{x}^k = 0, \tag{6.18}$$

where the symbol Γ^i_{jk} denotes the so–called *affine connection* which is the source of *curvature*, which is geometric description for *noise* (see [Ing97, Ing98]). The higher the local curvatures of the skin manifold–patch M_2, the greater the internal EMS–noise. This noise is the source of our micro–level fat–related fluctuations.

Assuming that the electro–physiological principles of the EMS–based body–shaping are identical (only less subtle) to the principles of the EMS–based face–shaping, in the following subsections, we will focus on the facial anatomical geometry.

Local Facial Curvatures and Their Deviations

In this subsection we consider human face, with its distinguished local anatomical features, as a 2D Riemannian manifold (i.e, a smooth skin manifold–patch M_2), determined by muscular structural and functional anatomy (see Figure 6.5). Here we demonstrate that this anatomical geometry is not static, but

Facial Musculature				
Muscle	**Origin**	**Insertion**	**Action**	**Innervation**
depressor anguli oris	oblique line of mandible	angle of mouth	pulls the corner of the mouth downward	marginal mandibular & buccal branches of facial nerve (VII)
levator labii superioris	inferior margin of orbit	skin of upper lip	elevates upper lip	buccal branch of facial nerve (VII)
zygomaticus major	zygomatic bone, upper lateral surface	skin of angle of mouth	elevates and draws the corner of mouth laterally	zygomatic and buccal branches of facial nerve (VII)
orbicularis oris	skin and fascia of lips and area surrounding lips	skin and fascia of lips	purses the lips	buccal branch of facial nerve (VII)
buccinator	pterygomandibular raphe, mandible, and maxilla lateral to molar teeth	angle of mouth & lateral portion of upper and lower lips	pulls corner of mouth laterally; pulls cheek against teeth	buccal branches of facial nerve (VII)
platysma	fascia overlying the pectoralis major and deltoid muscles	inferior border of mandible and skin of lower face	draws corners of mouth down; aids in depression of the mandible	facial nerve (VII), cervical branch

Table 6.1. Functional anatomy of the facial musculature (for whole body anatomy, see any anatomical textbook, e.g., [Mar98].)

rather dynamic in an extremely complex way, which can be controlled by a proper EMS.

Recall from Chapter 2, that Riemannian *metric* on any smooth nD Riemannian manifold M is a positive–definite quadratic form $g : M \to \mathbb{R}$, which is in local coordinates $(x^1(s), ..., x^n(s))$, dependent on the affine line parameter s at a point $m \in M$, defined as a symmetric $(0, 2)$ tensor–field $g_{ij}(m) = g\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right)(m)$, (see also [Boo86, Iva04, Iva02, IP01b, Iva05]).

An infinitesimal distance between the two nearby local points m and n on M is defined by the *line element*

$$ds^2 = g_{ij}\, dx^i dx^j,$$

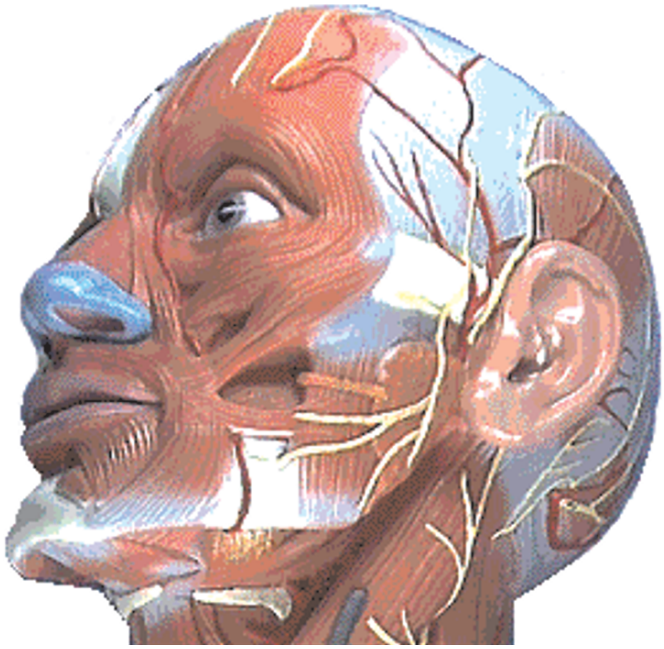

Fig. 6.5. Structural anatomy of the facial musculature (together with superficial branches of the facial nerve), showing local geometric features: distances (metrics), curvatures and directions for muscular contractions (see Table 6.1), modelled here as geodesic deviations. Here we show only facial EMS, assuming that all biophysical and geometric principles are the same (only less subtle) for body EMS.

and realized by the geodesics $x^i(s)$ (see Chapter 2). In local coordinates $(x^1(s), ..., x^n(s))$ at a point $m \in M$, the geodesic defining equation (6.18), derived from the *geometric action principle* (6.17), is a second order ordinary differential equation with the *Christoffel symbols* $\Gamma^i_{jk} = \Gamma^i_{jk}(m)$ of the *affine* (Levi–Civita) *connection* $\nabla_{\dot{x}^i}$ are calculated at the point $m \in M$ with local coordinates $(x^1(s), ..., x^n(s))$.

If $\Gamma^i_{jk}(m) = 0$, the manifold M is *flat* at the point m. This means that the *Riemann curvature tensor*, a symmetric $(1, 3)$ tensor field R^i_{jkl}=$R^i_{jkl}(m)$, locally defined at a point $m \in M$ as

$$R^i_{jkl} = \partial_{x^k}\Gamma^i_{jl} - \partial_{x^l}\Gamma^i_{jk} + \Gamma^i_{\mu k}\Gamma^\mu_{jl} - \Gamma^i_{\mu l}\Gamma^\mu_{jk},$$

also vanishes at a point $m \in M$, i.e., $R^i_{jkl}(m) = 0$.

Elliptic manifolds have positive curvature, i.e., $R^i_{jkl}(m) > 0$ and nearby geodesics are converging on it (see Figure 6.5), while hyperbolic ones have negative curvature, i.e., $R^i_{jkl}(m) < 0$ and nearby geodesics are diverging on it (see Figure 6.6). In case of a surface, which is a 2D manifold, the metric tensor is reduced to the scalar curvature R. An example of elliptic (convex)

surfaces is the sphere with the curvature $R = +\rho^{-2}$ (where ρ is the radius), while an example of hyperbolic (concave) surfaces is the Lobachevsky plane with the curvature $R = -1$.

Now, recall from section 2.5 above, that the rate of change of a vector field A^k on the manifold M along the curve $x^i(s)$ is properly defined by the *covariant derivative*:

$$\frac{D}{ds}A^k = \dot{x}^i \nabla_{\dot{x}^i} A^k = \dot{x}^i \left(\partial_{x^i} A^k + \Gamma^k_{ij} A^j\right) = \dot{A}^k + \Gamma^k{}_{ij}\, \dot{x}^i A^j,$$

which defines the parallel transport along the curve $x^i(s)$ at a point $m \in M$ as $\frac{D}{ds}A^k(m) = 0$.

By applying the previous result to itself, we can obtain an expression for the second covariant derivative of the vector field A^k along the curve $x^i(s)$:

$$\frac{D^2}{ds^2}A^k = \frac{d}{ds}\left(\dot{A}^k + \Gamma^k{}_{ij}\, \dot{x}^i A^j\right) + \Gamma^k{}_{ij}\, \dot{x}^i \left(\dot{A}^j + \Gamma^j{}_{mn}\, \dot{x}^m A^n\right).$$

In the same local coordinates $(x^1(s), ..., x^n(s))$ at a point $m \in M$, let $\delta x^i = \delta x^i(s)$ denote the vector–field of *geodesic deviation*, i.e., the infinitesimal vector–field describing both *normal* and *tangential* separation between the two neighboring geodesics, then the *Jacobi equation* of the geodesic deviation on the manifold M holds ([Arn89]):

$$\frac{D^2\delta x^i}{ds^2} + R^i_{jkl}\, \dot{x}^j\, \delta x^k\, \dot{x}^l = 0. \tag{6.19}$$

This equation describes the *relative acceleration* between two infinitesimally close facial geodesics, which is proportional both to the facial curvature (measured by the Riemann tensor R^i_{jkl} at a point $m \in M$), and to the geodesic deviation δx^i. Solutions of the Jacobi equation (6.19) are called *Jacobi fields*, or *Jacobi flows*.

Human Face as a Riemannian Patch–Manifold

Local anatomical features of the human face can be considered as a collection of local 2D Riemannian manifold–patches M^i_2, i.e., Riemannian patches, determined by skeletal and muscular anatomy. Each of these local patches M^i_2 is represented by its own local coordinates $(x^1(s), x^2(s))_i$, defined at a distinguished point m. We propose here a 2D Jacobi fields $\delta x^1, \delta x^2$ (see Figures 6.5 and 6.6) to model respectively normal and tangential components of contractions of the facial muscles.

In $2D$, the Riemann curvature tensor simplifies into:

$$R^i_{jmn} = \frac{1}{2} R\, g^{ik}(g_{km}\, g_{jn} - g_{kn}\, g_{jm}),$$

where R denotes the *scalar curvature*. Consequently the equation of geodesic deviation (6.19) also simplifies into

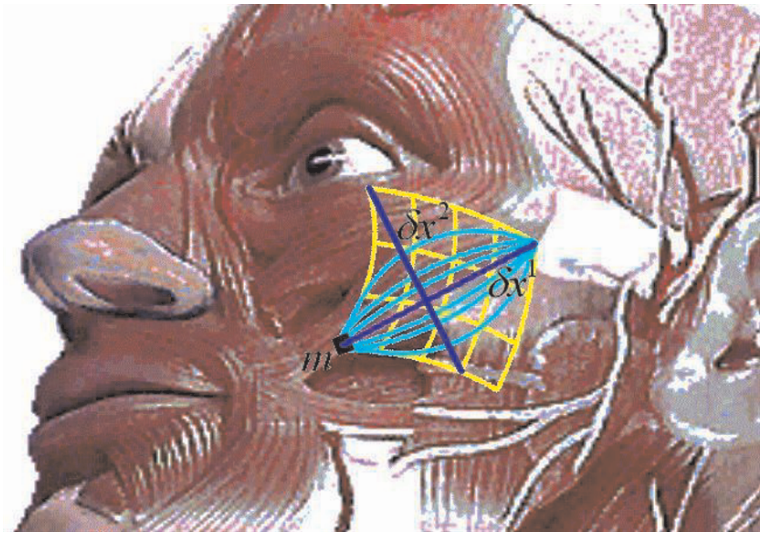

Fig. 6.6. Local coordinate chart defined in a neighborhood of a point m on the *convex–ecliptic* patch–manifold M_2 of the facial musculature, together with the converging geodesic deviation: its tangent component δx^1 and its normal component δx^2.

$$\frac{D^2}{ds^2}\delta x^i + \frac{R}{2}\delta x^i - \frac{R}{2}\dot{x}^i(g_{jk}\,\dot{x}^j\,\delta x^k) = 0. \tag{6.20}$$

Now, if we work in a local Cartesian coordinate system, defined at the tangent plane T_mM_2 at a point m by an orthogonal projection imaging, the covariant derivative $\frac{D^2}{ds^2}$ reduces to the ordinary derivative $\frac{d^2}{ds^2}$ (as the Christoffel symbols Γ^i_{jk} vanish) and the metric tensor g_{ij} reduces to identity matrix I_{ij}, so our $2D$ equation of geodesic deviation (6.20) reduces into a simple second order ordinary differential equation in just two coordinates $x^i\,(i = 1, 2)$

$$\frac{d^2}{ds^2}\delta x^i + \frac{R}{2}\delta x^i - \frac{R}{2}\dot{x}^i(I_{jk}\,\dot{x}^j\,\delta x^k) = 0.$$

Also, if we require that the two nearby geodesics be nearly parallel, the last term in (6.20) vanishes, and we are left with

$$\frac{D^2}{ds^2}\delta x^i + \frac{R}{2}\delta x^i = 0. \tag{6.21}$$

Again, if we work in a locally Cartesian coordinate system, our *flat* $2D$ *equation of geodesic deviation* simplifies into *harmonic oscillator* in which the scalar curvature $R/2$ plays the role of the *spring constant*:

$$\frac{d^2}{ds^2}\delta x^i + \frac{R}{2}\delta x^i = 0. \tag{6.22}$$

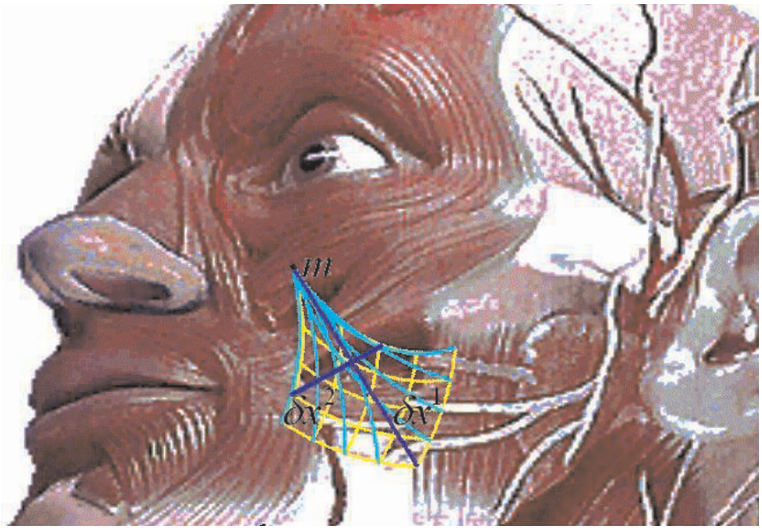

Fig. 6.7. Local coordinate chart defined in a neighborhood of a point m on the *concave–hyperbolic* patch–manifold M_2 of the facial musculature, together with the diverging geodesic deviation: its tangent component δx^1 and its normal component δx^2.

Therefore, equations (6.21) and (6.20) could be respectively regarded as the first–order and second–order perturbations of the linear oscillator equation (6.22). These three equations represent the three levels of detail in our modelling of the facial muscular movements. The oscillator equation (6.22) has a simple family of sinus functions (with certain amplitudes, frequencies and phases) as a solution, while the two nonlinear equations (6.21) and (6.20) could be numerically integrated for zero initial deviation and its velocity, using any explicit Runge–Kutta–like integrator (see, e.g. [IS01]). Each of them describes the facial movement caused by muscular contraction dependent on its local curvature, i.e., anatomical shape. Also, all three geometric oscillators have kinetic and potential energies respectively defined as quadratic forms:

$$E_{kin} = \frac{1}{2} g_{ij} \delta\dot{x}^i \delta\dot{x}^j, \qquad U = \frac{1}{4} R g_{ij} \delta x^i \delta x^j,$$

and derived from the muscular action principle (6.12–6.13) above.

6.5.2 Cellular Muscle–Fat Geometry

On the micro–level in the phase EMS_{geom}, we have an adaptive *sum over fractal geometries*, represented by the path integral over all regional Riemannian metrics $g_{ij} = g_{ij}(x)$ varying from point to point inside an n–dimensional *muscle–fat manifold* M, underlying the external skin surface,

$$SHAPE_{geom} = \int \mathcal{D}[wg_{ij}]\, \mathrm{e}^{\mathrm{i}S[g_{ij}]} \quad \underrightarrow{Wick} \quad \int \mathcal{D}[wg_{ij}]\, \mathrm{e}^{-S[g_{ij}]}, \tag{6.23}$$

where $\mathcal{D}[g_{ij}]$ denotes diffeomorphism equivalence classes of metrics $g_{ij}(x)$ of *Skin*.

To include the severe change of topological structure (e.g., a change in a number of holes) in the manifold M, equation (6.23) can be extended as

$$SHAPE_{geom/top} = \sum_{\text{topol.}} \int \mathcal{D}[wg_{ij}]\, \mathrm{e}^{\mathrm{i}S[g_{ij}]}, \tag{6.24}$$

where the topological sum is taken over all components of connectedness of the manifold M determined by its *Euler characteristics* [Iva02]. This type of integral defines the *theory of fluctuating geometries*, a propagator between $(n-1)-$dimensional boundaries of the $n-$dimensional manifold M. One has to contribute a meaning to the integration over geometries. A key ingredient in doing so is to approximate in a natural way the smooth structures of the manifold M by piecewise linear structures (mostly using topological simplices Δ [8]). In this way, after the Wick–rotation (6.15), the integral (6.23–6.24) becomes a *simple statistical system*, given by partition function

$$Z = \sum_{\Delta} \frac{1}{C_{\Delta}} e^{-S_{\Delta}},$$

where the summation is over all triangulations Δ of the manifold M, while the number C_T is the order of the automorphism group of the performed triangulation.

[8] This is called the *simplicial approximation*.

A

Appendix

A.1 Basic Formulas from Tensor Analysis

Biomechanical laws must be independent of any particular coordinate systems used in describing them mathematically, if they are to be valid. In other words, all biomechanical equations need to be tensorial or *covariant*. Therefore, for the reference purpose, in this subsection, we give the basic formulas from the standard tensor calculus, which is used throughout the text. The basic notational convention used in tensor calculus is Einstein's summation convention over repeated indices. More on this subject can be found in any standard textbook on mathematical methods for scientists and engineers, or mathematical physics (we recommend [MTW73]).

A.1.1 Transformation of Coordinates and Elementary Tensors

To introduce tensors, consider a standard linear nD matrix system, $\mathbf{Ax} = \mathbf{b}$. It can be rewritten in the so–called *covariant form* as

$$a_{ij}x^j = b_i\,, \qquad (i,j = 1,...,n). \tag{A.1}$$

Here, i is a *free index* and j is a *dummy index* to be summed upon, so the expansion of (A.1) gives

$$\begin{aligned} a_{11}x^1 + a_{12}x^2 + ... + a_{1n}x^n &= b_1\,,\\ a_{21}x^1 + a_{22}x^2 + ... + a_{2n}x^n &= b_2\,,\\ &... \\ a_{n1}x^1 + a_{n2}x^2 + ... + a_{nn}x^n &= b_n\,, \end{aligned}$$

as expected from the original matrix form $\mathbf{Ax} = \mathbf{b}$. This indicial notation can be more useful than the matrix one, like e.g., in computer science, where indices would represent loop variables. However, the full potential of tensor analysis is to deal with nonlinear multivariate systems, which are untractable by linear matrix algebra and analysis. The core of this *nonlinear multivariate analysis* is *general functional transformation*.

Transformation of Coordinates

Suppose that we have two sets of curvilinear coordinates that are single–valued, continuous and smooth functions of time, $x^j = x^j(t),\ (j = 1, ..., m)$ and $\bar{x}^i = \bar{x}^i(t),\ (i = 1, ..., n)$, respectively, representing trajectories of motion of some biomechanical system. Then a general $(m \times n)$D transformation (i.e., a nonlinear map) $x^j \mapsto \bar{x}^i$ is defined by the set of transformation equations

$$\bar{x}^i = \bar{x}^i(x^j), \qquad (i = 1, ..., n;\ j = 1, ..., m). \tag{A.2}$$

In case of the square transformation, $m = n$, we can freely exchange the indices, like e.g., in general relativity theory. On the other hand, in the general case of rectangular transformation, $m \neq n$, like e.g., in robotics, and we need to take care of these 'free' indices.

Now, if the *Jacobian determinant* of this *coordinate transformation* is different from zero,

$$\left| \frac{\partial \bar{x}^i}{\partial x^j} \right| \neq 0,$$

then the transformation (A.2) is reversible and the inverse transformation,

$$x^j = x^j(\bar{x}^i),$$

exists as well. Finding the inverse transformation is the problem of matrix inverse: in case of the square matrix it is well defined, although the inverse might not exist if the matrix is singular. However, in case of the square matrix, its proper inverse does not exist, and the only tool that we are left with is the so–called *Moore–Penrose pseudoinverse*, which gives an optimal solution (in the least–squares sense) of an overdetermined system of equations. Every (overdetermined) rectangular coordinate transformation gives rise to a *redundant system*.

For example, in Euclidean 3D space $\mathbb{R}^3$, transformation from Cartesian coordinates $y^k = \{x, y, z\}$ into spherical coordinates $x^i = \{\rho, \theta, \varphi\}$ is given by

$$y^1 = x^1 \cos x^2 \cos x^3, \qquad y^2 = x^1 \sin x^2 \cos x^3, \qquad y^3 = x^1 \sin x^3, \tag{A.3}$$

with the Jacobian matrix given by

$$\left(\frac{\partial y^k}{\partial x^i} \right) = \begin{pmatrix} \cos x^2 \cos x^3 & -x^1 \sin x^2 \cos x^3 & -x^1 \cos x^2 \sin x^3 \\ \sin x^2 \cos x^3 & x^1 \cos x^2 \cos x^3 & -x^1 \sin x^2 \sin x^3 \\ \sin x^3 & 0 & x^1 \cos x^3 \end{pmatrix} \tag{A.4}$$

and the corresponding Jacobian determinant, $\left| \frac{\partial y^k}{\partial x^i} \right| = (x^1)^2 \cos x^3$.

An inverse transform is given by

$$x^1 = \sqrt{(y^1)^2 + (y^2)^2 + (y^3)^2}, \qquad x^2 = \arctan\left(\frac{y^2}{y^1} \right),$$

$$x^3 = \arctan\left(\frac{y^3}{\sqrt{(y^1)^2 + (y^2)^2}} \right), \qquad \text{with} \quad \left| \frac{\partial x^i}{\partial y^k} \right| = \frac{1}{(x^1)^2 \cos x^3}.$$

As a main biomechanical example, we have a rectangular transformation from 6 DOF external, end–effector (e.g., hand) coordinates, into n DOF internal, joint–angle coordinates. In most cases this is a redundant manipulator system, with infinite number of possible joint trajectories.

Scalar Invariants

A *scalar invariant* (or, a zeroth order tensor) with respect to the transformation (A.2) is the quantity $\varphi = \varphi(t)$ defined as

$$\varphi(x^i) = \bar{\varphi}(\bar{x}^i),$$

which does not change at all under the coordinate transformation. In other words, φ is *invariant* under (A.2). Biodynamic examples of scalar invariants include various energies (kinetic, potential, biochemical, mental) with the corresponding kinds of work, as well as related thermodynamic quantities (free energy, temperature, entropy, etc.).

Vectors and Covectors

Any geometric object $v^i = v^i(t)$ that under the coordinate transformation (A.2) transforms as

$$\bar{v}^i = v^j \frac{\partial \bar{x}^i}{\partial x^j}, \qquad \text{(remember, summing upon } j-\text{index)},$$

represents a *vector*, traditionally called a *contravariant vector*, or, a first–order contravariant tensor. Standard biomechanical examples include both translational and rotational velocities and accelerations.

On the other hand, any geometric object $v_i = v_i(t)$ that under the coordinate transformation (A.2) transforms as

$$\bar{v}_i = v_j \frac{\partial x^j}{\partial \bar{x}^i},$$

represents a *one–form* or *covector*, traditionally called a *covariant vector*, or, a first order covariant tensor. Standard biomechanical examples include both translational and rotational momenta, forces and torques.

Second–Order Tensors

Any geometric object $t^{ik} = t^{ik}(t)$ that under the coordinate transformation (A.2) transforms as

$$\bar{t}^{ik} = t^{jl} \frac{\partial \bar{x}^i}{\partial x^j} \frac{\partial \bar{x}^k}{\partial x^l}, \qquad (i,k = 1,...,n;\ j,l = 1,...,m),$$

represents a *second–order contravariant tensor*. It can be obtained as an *outer product* of two contravariant vectors, $t^{ik} = u^i v^k$.

Any geometric object $t_{ik} = t_{ik}(t)$ that under the coordinate transformation (A.2) transforms as

$$\bar{t}_{ik} = t_{jl} \frac{\partial x^j}{\partial \bar{x}^i} \frac{\partial x^l}{\partial \bar{x}^k},$$

represents a *second–order covariant tensor*. It can be obtained as an outer product of two covariant vectors, $t_{ik} = u_i v_k$.

Any geometric object $t^i_k = t^i_k(t)$ that under the coordinate transformation (A.2) transforms as

$$\bar{t}^i_k = t^j_l \frac{\partial \bar{x}^i}{\partial x^j} \frac{\partial x^l}{\partial \bar{x}^k},$$

represents a *second–order mixed tensor*. It can be obtained as an outer product of a covariant vector and a contravariant vector, $t^i_k = u^i v_k$.

Standard biomechanical examples include:

1. The fundamental (material) covariant metric tensor $\mathbf{g} \equiv g_{ik}$, i.e., inertia matrix, given usually by the transformation from Cartesian coordinates y^j to curvilinear coordinates x^i,

   $$g_{ik} = \frac{\partial y^j}{\partial x^i} \frac{\partial y^j}{\partial x^k}, \qquad \text{(summing over } j\text{)}.$$

   It is used in the quadratic metric form ds^2 of the space in consideration (e.g., a certain biomechanical configuration space)

   $$ds^2 \equiv dy^j dy^j = g_{ik} dx^i dx^k,$$

   where the first term on the r.h.s denotes the *Euclidean metrics*, while the second term is the *Riemannian metric* of the space, respectively.
2. Its inverse $\mathbf{g}^{-1} \equiv g^{ik}$, given by

   $$g^{ik} = (g_{ik})^{-1} = \frac{G_{ik}}{|g_{ik}|}, \qquad G_{ik} \text{ is the cofactor of the matrix } (g_{ik});$$

3. The *Kronecker–delta* symbol δ^i_k, given by

   $$\delta^i_k = \begin{cases} 1 \text{ if } i = k \\ 0 \text{ if } i \neq k \end{cases},$$

   used to denote the metric tensor in Cartesian orthogonal coordinates. δ^i_k is a discrete version of the *Dirac* $\delta-$*function*. The *generalized Kronecker–delta symbol* δ^{ijk}_{lmn} (in 3D) is the product of *Ricci antisymmetric tensors* ε^{ijk} and ε_{lmn},

   $$\delta^{ijk}_{lmn} = \varepsilon^{ijk}\varepsilon_{lmn} = \begin{cases} 0 & \text{if at least two indices are equal} \\ +1 & \text{if both } ijk \text{ and } lmn \text{ are either even or odd} \\ -1 & \text{if one of } ijk,\ lmn \text{ is even and the other is odd} \end{cases}.$$

For example, to derive components of the metric tensor $\mathbf{g} \equiv g_{ij}$ in standard spherical coordinates, we use the relations (A.3–A.4) between the spherical coordinates $x^i = \{\rho, \theta, \varphi\}$ and the Cartesian coordinates $y^k = \{x, y, z\}$, and the definition, $g_{ij} = \frac{\partial y^k}{\partial x^i} \frac{\partial y^k}{\partial x^j}$, to get the metric tensor (in matrix form)

$$(g_{ij}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & (x^1)^2 \cos^2 x^3 & 0 \\ 0 & 0 & (x^1)^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \rho^2 \cos^2 \varphi & 0 \\ 0 & 0 & \rho^2 \end{pmatrix}, \tag{A.5}$$

and the inverse metric tensor

$$(g^{ij}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{(x^1)^2 \cos^2 x^3} & 0 \\ 0 & 0 & \frac{1}{(x^1)^2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\rho^2 \cos^2 \varphi} & 0 \\ 0 & 0 & \frac{1}{\rho^2} \end{pmatrix}. \tag{A.6}$$

Given a tensor, we can derive other tensors by raising and lowering its indices, by their multiplication with covariant and contravariant metric tensors. In this way, the so–called *associated tensors* to the given tensor are be formed. For example, v^i and v_i are associated tensors, related by

$$v_i = g_{ik} v^k \qquad \text{and} \qquad v^i = g^{ik} v_k.$$

Given two vectors, $\mathbf{u} \equiv u^i$ and $\mathbf{v} \equiv v^i$, their inner (dot, or scalar) product is given by

$$\mathbf{u} \cdot \mathbf{v} \equiv g_{ij} u^i v^j,$$

while their vector (cross) product (in 3D) is given by

$$\mathbf{u} \times \mathbf{v} \equiv \varepsilon_{ijk} u^j v^k.$$

Higher–Order Tensors

As a generalization of above tensors, consider a geometric object $R^i_{kps} = R^i_{kps}(t)$ that under the coordinate transformation (A.2) transforms as

$$\bar{R}^i_{kps} = R^j_{lqt} \frac{\partial \bar{x}^i}{\partial x^j} \frac{\partial x^l}{\partial \bar{x}^k} \frac{\partial x^q}{\partial \bar{x}^p} \frac{\partial x^t}{\partial \bar{x}^s}, \qquad (\text{all indices} = 1, ..., n). \tag{A.7}$$

Clearly, $R^i_{kjl} = R^i_{kjl}(x, t)$ is a fourth order tensor, once contravariant and three times covariant, representing the central tensor in Riemannian geometry, called the *Riemann curvature tensor*. As all biomechanical configuration spaces are Riemannian manifolds, they are all characterized by curvature tensors. In case $R^i_{kjl} = 0$, the corresponding Riemannian manifold reduces to the Euclidean space of the same dimension, in which $g_{ik} = \delta^i_k$.

If one contravariant and one covariant index of a tensor a set equal, the resulting sum is a tensor of rank two less than that of the original tensor. This process is called *tensor contraction*.

If to each point of a region in an nD space there corresponds a definite tensor, we say that a *tensor–field* has been defined. In particular, this is a *vector–field* or a *scalar–field* according as the tensor is of rank one or zero. It should be noted that a tensor or tensor field is not just the set of its components in one special coordinate system, but all the possible sets of components under any transformation of coordinates.

Tensor Symmetry

A tensor is called *symmetric* with respect to two indices of the same variance if its components remain unaltered upon interchange of the indices; e.g., $a_{ij} = a_{ji}$, or $a^{ij} = a^{ji}$. A tensor is called *skew–symmetric* (or, *antisymmetric*) with respect to two indices of the same variance if its components change sign upon interchange of the indices; e.g., $a_{ij} = -a_{ji}$, or $a^{ij} = -a^{ji}$. Regarding tensor symmetry, in the following we will prove several useful propositions.

(i) *Every second–order tensor can be expressed as the sum of two tensors, one of which is symmetric and the other is skew–symmetric.* For example, a second order tensor a_{ij}, which is for $i, j = 1, ..., n$ given by the $n \times n-$matrix

$$a_{ij} = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{n2} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix},$$

can be rewritten as

$$\begin{aligned} a_{ij} &= \frac{1}{2}a_{ij} + \frac{1}{2}a_{ij} + \frac{1}{2}a_{ji} - \frac{1}{2}a_{ji}\,, \quad \text{that can be rearranged as} \\ &= \frac{1}{2}a_{ij} + \frac{1}{2}a_{ji} + \frac{1}{2}a_{ij} - \frac{1}{2}a_{ji}\,, \quad \text{which can be regrouped as} \\ &= \frac{1}{2}(a_{ij} + a_{ji}) + \frac{1}{2}(a_{ij} - a_{ji}), \quad \text{which can be written as} \\ &= a_{(ij)} + a_{[ij]}\,, \end{aligned}$$

where $a_{(ij)}$ denotes its symmetric part, while $a_{[ij]}$ denotes its skew–symmetric part, as required.

(ii) *Every quadratic form can be made symmetric.* For example, a quadratic form $a_{ij}x^i x^j$, that (for $i, j = 1, ..., n$) expands as

$$\begin{aligned} a_{ij}x^i x^j = {} & a_{11}x^1x^1 + a_{12}x^1x^2 + ... + a_{1n}x^1x^n + \\ & + a_{21}x^2x^1 + a_{22}x^2x^2 + ... + a_{2n}x^2x^n + \\ & ... \\ & + a_{n1}x^nx^1 + a_{n2}x^nx^2 + ... + a_{nn}x^nx^n, \end{aligned}$$

with a non–symmetric second order tensor a_{ij}, can be made symmetric in the following way.

$$
\begin{aligned}
a_{ij}x^i x^j &= \frac{1}{2}a_{ij}x^i x^j + \frac{1}{2}a_{ij}x^i x^j. \\
&\quad \text{If we swap indices in the second term, we get} \\
&= \frac{1}{2}a_{ij}x^i x^j + \frac{1}{2}a_{ji}x^j x^i\,, \quad \text{which is equal to} \\
&= \frac{1}{2}(a_{ij} + a_{ji})\, x^i x^j. \\
&\quad \text{If we now use a substitution,} \\
\frac{1}{2}(a_{ij} + a_{ji}) \equiv b_{ij} &= b_{ji}, \quad \text{we get} \\
a_{ij}x^i x^j &= b_{ij}x^i x^j,
\end{aligned}
$$

where a_{ij} is non–symmetric and b_{ij} is symmetric, as required.

(iii) *Every second order tensor that is the sum* $a^{ij} = u^i v^j + u^j v^i$, or, $a_{ij} = u_i v_j + u_j v_i$ is *symmetric.* In both cases, if we swap the indices i and j, we get $a^{ji} = u^j v^i + u^i v^j$, (resp. $a_{ji} = u_j v_i + u_i v_j$), which implies that the tensor a^{ij} (resp. a_{ij}) is symmetric.

(iv) *Every second order tensor that is the difference* $b^{ij} = u^i v^j - u^j v^i$, or, $b_{ij} = u_i v_j - u_j v_i$ is *skew–symmetric.* In both cases, if we swap the indices i and j, we get $b^{ji} = -(u^j v^i - u^i v^j)$, (resp. $b_{ji} = -(u_j v_i - u_i v_j)$), which implies that the tensor b^{ij} (resp. b_{ij}) is skew–symmetric.

A.1.2 Euclidean Tensors

Basis Vectors and the Metric Tensor in $\mathbb{R}^n$

The natural *Cartesian coordinate basis* in an nD Euclidean space $\mathbb{R}^n$ is defined as a set of nD unit vectors e^i given by

$$e^1 = [\{1,0,0,...\}^t,\ e^2 = \{0,1,0,...\}^t,\ e^3 = \{0,0,1,...\}^t, ...,\ e^n = \{0,0,...,1\}^t],$$

(where index t denotes transpose) while its dual basis e_i is given by:

$$e_1 = [\{1,0,0,...\},\ e_2 = \{0,1,0,...\},\ e_3 = \{0,0,1,...\}, ...,\ e_n = \{0,0,...,1\}],$$

(no transpose) where the definition of the dual basis is given by the Kronecker's $\delta-$symbol, i.e., the $n \times n$ identity matrix:

$$e^i \cdot e_j = \delta^i_j = \begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 1 \end{bmatrix},$$

that is the metric tensor in Cartesian coordinates equals $\mathbf{g} = \delta^i_j$. In general, (i.e., curvilinear) coordinate system, the metric tensor $\mathbf{g} = g_{ij}$ is defined as the scalar product of the dual basis vectors, i.e., the $n \times n$ matrix:

$$g_{ij} = e_i \cdot e_j = \begin{bmatrix} g_{11} & g_{12} & g_{13} & \cdots & g_{1n} \\ g_{21} & g_{22} & g_{23} & \cdots & g_{2n} \\ g_{31} & g_{32} & g_{33} & \cdots & g_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ g_{n1} & g_{n2} & g_{n3} & \cdots & g_{nn} \end{bmatrix}.$$

Tensor Products in $\mathbb{R}^n$

Let $\mathbf{u}$ and $\mathbf{v}$ denote two vectors in $\mathbb{R}^n$, with their components given by

$$u^i = u \cdot e^i, \quad \text{and} \quad v^j = v \cdot e^j,$$

where $u = |\mathbf{u}|$ and $v = |\mathbf{v}|$ are their respective norms (or, lengths). Then their inner product (i.e., scalar, or dot product) $\mathbf{u} \cdot \mathbf{v}$ is a scalar invariant S, defined as

$$S = u^i \cdot v^j = g_{ij} u^i v^j.$$

Besides the dot product of two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, there is also their tensor product (i.e., generalized vector, or cross product), which is a second order tensor

$$\mathbf{T} = \mathbf{u} \otimes \mathbf{v}, \text{ in components, } T^{ij} = u^i \otimes v^j.$$

In the natural basis e_i this tensor is expanded as

$$\mathbf{T} = T^{ij} e_i \otimes e_j,$$

while its components in the dual basis read:

$$T^{ij} = T(e^i, e^j),$$

where $T = |\mathbf{T}|$ is its norm. To get its components in curvilinear coordinates, we need first to substitute it in Cartesian basis:

$$T^{ij} = T^{mn}(e_m \otimes e_n)(e^i, e^j),$$

then to evaluate it on the slots:

$$T^{ij} = T^{mn} e_m \cdot e^i \, e_n \cdot e^j,$$

and finally to calculate the other index configurations by lowering indices, by means of the metric tensor:

$$T^i_j = g_{jm} T^{im}, \qquad T_{ij} = g_{im} g_{jn} T^{mn}.$$

A.1.3 Tensor Derivatives on Riemannian Manifolds

Consider now some nD Riemannian manifold M with the metric form (i.e., line element) $ds^2 = g_{ik} dx^i dx^k$, as a configuration space for a certain biomechanical system (e.g., human spine, or arm–shoulder complex).

Christoffel's Symbols

Partial derivatives of the metric tensor g_{ik} form themselves special symbols that do not transform as tensors (with respect to the coordinate transformation (A.2)), but nevertheless represent important quantities in tensor analysis. They are called *Christoffel symbols of the first kind*, defined by

$$\Gamma_{ijk} = \frac{1}{2}(\partial_{x^i} g_{jk} - \partial_{x^j} g_{ki} + \partial_{x^k} g_{ij}), \qquad \left(\text{remember, } \partial_{x^i} \equiv \frac{\partial}{\partial x^i}\right)$$

and *Christoffel symbols of the second kind*, defined by

$$\Gamma^k_{ij} = g^{kl}\Gamma_{ijl}.$$

The Riemann curvature tensor R^l_{ijk} (A.7) of the manifold M, can be expressed in terms of the later as

$$R^l_{ijk} = \partial_{x^j}\Gamma^l_{ik} - \partial_{x^k}\Gamma^l_{ij} + \Gamma^l_{rj}\Gamma^r_{ik} - \Gamma^l_{rk}\Gamma^r_{ij}.$$

For example, in 3D spherical coordinates, $x^i = \{\rho, \theta, \varphi\}$, with the metric tensor and its inverse given by (A.5, A.6), it can be shown that the only nonzero Christoffel's symbols are:

$$\begin{aligned} &\Gamma^2_{12} = \Gamma^2_{21} = \Gamma^3_{13} = \Gamma^3_{31} = \frac{1}{\rho}, \qquad \Gamma^3_{23} = \Gamma^2_{32} = -\tan\theta, \\ &\Gamma^1_{22} = -\rho, \qquad \Gamma^1_{33} = -\rho\cos^2\theta, \qquad \Gamma^2_{33} = \sin\theta\cos\theta. \end{aligned} \tag{A.8}$$

Geodesics

From the Riemannian metric form $ds^2 = g_{ik}dx^i dx^k$ it follows that the distance between two points t_1 and t_2 on a curve $x^i = x^i(t)$ in M is given by

$$s = \int_{t_1}^{t_2} \sqrt{g_{ik}\dot{x}^i\dot{x}^k}dt.$$

That curve $x^i = x^i(t)$ in M which makes the distance s a minimum is called a *geodesic* of the space M (e.g., in a sphere, the geodesics are arcs of great circles). Using the calculus of variations, the geodesics are found from the differential *geodesic equation*,

$$\ddot{x}^i + \Gamma^i_{jk}\dot{x}^j\dot{x}^k = 0, \tag{A.9}$$

where overdot means derivative upon the line parameter s.

For example, in 3D spherical coordinates $x^i = \{\rho, \theta, \varphi\}$, using (A.8), geodesic equation (A.9) becomes a system of three scalar ODEs,

$$\begin{aligned} &\ddot{\rho} - \rho\dot{\theta}^2 - \rho\cos^2\theta\dot{\varphi}^2 = 0, \qquad \ddot{\theta} + \frac{2}{\rho}\dot{\rho}\dot{\varphi} + \sin\theta\cos\theta\dot{\varphi}^2 = 0, \\ &\ddot{\varphi} + \frac{2}{\rho}\dot{\rho}\dot{\varphi} - 2\tan\theta\dot{\theta}\dot{\varphi} = 0. \end{aligned} \tag{A.10}$$

The Covariant Derivative

Ordinary total and partial derivatives of vectors (covectors) *do not transform as vectors* (covectors) with respect to the coordinate transformation (A.2). For example, let y^k be Cartesian coordinates and x^i be general curvilinear coordinates of a dynamical system (with $i,k = 1,...,n$). We have: $x^i(t) = x^i[y^k(t)]$, which implies that

$$\frac{dx^i}{dt} = \frac{\partial x^i}{\partial y^k}\frac{dy^k}{dt}, \qquad \text{or equivalently,} \qquad \dot{x}^i = \frac{\partial x^i}{\partial y^k}\dot{y}^k,$$

that is a transformation law for the contravariant vector, which means that the velocity $v^i \equiv \dot{x}^i \equiv \frac{dx^i}{dt}$ *is* a proper contravariant vector. However, if we perform another time differentiation, we get

$$\frac{d^2x^i}{dt^2} = \frac{\partial x^i}{\partial y^k}\frac{d^2y^k}{dt^2} + \frac{\partial^2 x^i}{\partial y^k \partial y^m}\frac{dy^k}{dt}\frac{dy^m}{dt},$$

which means that $\frac{d^2x^i}{dt^2}$ *is not* a proper vector.

$\frac{d^2x^i}{dt^2}$ is an acceleration vector only in a special case when x^i are another Cartesian coordinates; then $\frac{\partial^2 x^i}{\partial y^k \partial y^m} = 0$, and therefore the original coordinate transformation is linear, $x^i = a^i_k y^k + b^i$ (where a^i_k and b^i are constant).

Therefore, $\frac{d^2x^i}{dt^2}$ represents an acceleration vector only in terms of Newtonian mechanics in a Euclidean space $\mathbb{R}^n$, while it is not a proper acceleration vector in terms of Lagrangian or Hamiltonian mechanics in general curvilinear coordinates on a smooth manifold M^n. And we know that Newtonian mechanics in $\mathbb{R}^n$ is sufficient only for fairly simple mechanical systems.

The above is true for any tensors. So we need to find another derivative operator to be able to preserve their tensor character. The solution to this problem is called the *covariant derivative*.

The covariant derivative $v^i_{;k}$ of a contravariant vector v^i is defined as

$$v^i_{;k} = \partial_{x^k} v^i + \Gamma^i_{jk} v^j.$$

Similarly, the covariant derivative $v_{i;k}$ of a covariant vector v_i is defined as

$$v_{i;k} = \partial_{x^k} v_i - \Gamma^j_{ik} v_j.$$

Generalization for the higher order tensors is straightforward; e.g., the covariant derivative $t^j_{kl;q}$ of the third order tensor t^j_{kl} is given by

$$t^j_{kl;q} = \partial_{x^q} t^j_{kl} + \Gamma^j_{qs} t^s_{kl} - \Gamma^s_{kq} t^j_{sl} - \Gamma^s_{lq} t^j_{ks}.$$

The covariant derivative is the most important tensor operator in general relativity (its zero defines *parallel transport*) as well as the basis for defining other differential operators in mechanics and physics.

Covariant Form of Gradient, Divergence, Curl and Laplacian

Gradient. If $\varphi = \varphi(x^i, t)$ is a scalar field, the gradient one–form $\mathrm{grad}(\varphi)$ is defined by

$$\mathrm{grad}(\varphi) = \nabla\varphi = \varphi_{;i} = \partial_{x^i}\varphi.$$

Divergence. The divergence $\mathrm{div}(v^i)$ of a vector–field $v^i = v^i(x^i, t)$ is defined by contraction of its covariant derivative with respect to the coordinates $x^i = x^i(t)$, i.e., the contraction of $v^i_{;k}$, namely

$$\mathrm{div}(v^i) = v^i_{;i} = \frac{1}{\sqrt{g}}\partial_{x^i}(\sqrt{g}v^i).$$

Curl. The curl $\mathrm{curl}(\theta_i)$ of a one–form $\theta_i = \theta_i(x^i, t)$ is a second order covariant tensor defined as

$$\mathrm{curl}(\theta_i) = \theta_{i;k} - \theta_{k;i} = \partial_{x^k}\theta_i - \partial_{x^i}\theta_k.$$

Laplacian. The Laplacian $\Delta\varphi$ of a scalar invariant $\varphi = \varphi(x^i, t)$ is the divergence of $\mathrm{grad}(\varphi)$, or

$$\Delta\varphi = \nabla^2\varphi = \mathrm{div}(\mathrm{grad}(\varphi)) = \mathrm{div}(\varphi_{;i}) = \frac{1}{\sqrt{g}}\partial_{x^i}(\sqrt{g}g^{ik}\partial_{x^k}\varphi).$$

The Absolute Derivative

The *absolute derivative* (or *intrinsic*, or *Bianchi's derivative*) of a contravariant vector v^i along a curve $x^k = x^k(t)$ is denoted by $\dot{\overset{\cdot}{v}}{}^i \equiv Dv^i/dt$ and defined as the inner product of the covariant derivative of v^i and $\dot{x}^k \equiv dx^k/dt$, i.e., $v^i_{;k}\dot{x}^k$, and is given by

$$\dot{\overset{\cdot}{v}}{}^i = \dot{v}^i + \Gamma^i_{jk}v^j\dot{x}^k.$$

Similarly, the absolute derivative $\dot{\overset{\cdot}{v}}_i$ of a covariant vector v_i is defined as

$$\dot{\overset{\cdot}{v}}_i = \dot{v}_i - \Gamma^j_{ik}v_j\dot{x}^k.$$

Generalization for the higher order tensors is straightforward; e.g., the absolute derivative $\dot{\overset{\cdot}{t}}{}^j_{kl}$ of the third order tensor t^j_{kl} is given by

$$\dot{\overset{\cdot}{t}}{}^j_{kl} = \dot{t}^j_{kl} + \Gamma^j_{qs}t^s_{kl}\dot{x}^q - \Gamma^s_{kq}t^j_{sl}\dot{x}^q - \Gamma^s_{lq}t^j_{ks}\dot{x}^q.$$

The absolute derivative is the most important operator in biomechanics, as it is the basis for the *covariant form* of both Lagrangian and Hamiltonian equations of motion of many biomechanical systems.

Application to Curve Geometry

Given three unit vectors: *tangent* τ^i, *principal normal* β^i, and *binormal* ν^i, as well as two scalar invariants: curvature K and torsion T, of a curve $\gamma(s) = \gamma[x^i(s)]$, the so–called *Frenet–Serret formulae* are valid[1]

$$\begin{aligned}
\dot{\tau}^i &\equiv \dot{\tau}^i + \Gamma^i_{jk}\tau^j\dot{x}^k = K\beta^i,\\
\dot{\beta}^i &\equiv \dot{\beta}^i + \Gamma^i_{jk}\beta^j\dot{x}^k = -(K\tau^i + T\nu^i),\\
\dot{\nu}^i &\equiv \dot{\nu}^i + \Gamma^i_{jk}\nu^j\dot{x}^k = T\beta^i.
\end{aligned}$$

Application to Mechanical Definitions of Acceleration and Force

In modern analytical mechanics, the two fundamental notions of *acceleration* and *force* in general curvilinear coordinates are substantially different from the corresponding terms in Cartesian coordinates as commonly used in engineering mechanics. Namely, 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. Proper mathematical definition of the acceleration vector is the absolute time derivative of the velocity vector, while the force is a differential one–form.

To give a brief look at these 'weird mathematical beasts', consider a material dynamical system described by n curvilinear coordinates $x^i = x^i(t)$. First, recall from subsection A.1.3 above, that an ordinary time derivative of the velocity vector $v^i(t) = \dot{x}^i(t)$ *does not transform as a vector* with respect to the general coordinate transformation (A.2). Therefore, $a^i \neq \dot{v}^i$. So, we need to use its absolute time derivative to define the acceleration vector (with $i, j, k = 1, ..., n$),

$$a^i = \dot{\bar{v}}^i \equiv \frac{Dv^i}{dt} = v^i_{;k}\dot{x}^k \equiv \dot{v}^i + \Gamma^i_{jk}v^jv^k \equiv \ddot{x}^i + \Gamma^i_{jk}\dot{x}^j\dot{x}^k, \tag{A.11}$$

which is equivalent to the l.h.s of the geodesic equation (A.9). Only in the particular case of Cartesian coordinates, the general acceleration vector (A.11) reduces to the familiar engineering form of the Euclidean acceleration vector[2], $\mathbf{a} = \dot{\mathbf{v}}$.

For example, in standard spherical coordinates $x^i = \{\rho, \theta, \varphi\}$, we have the components of the acceleration vector given by (A.10), if we now reinterpret overdot as the time derivative,

[1] In this paragraph, the overdot denotes the total derivative with respect to the line parameter s (instead of time t).

[2] Any Euclidean space can be defined as a set of Cartesian coordinates, while any Riemannian manifold can be defined as a set of curvilinear coordinates. Christoffel's symbols Γ^i_{jk} vanish in Euclidean spaces defined by Cartesian coordinates; however, they are nonzero in Riemannian manifolds defined by curvilinear coordinates (see Chapter 2 for geometric details).

$$a^{\rho} = \ddot{\rho} - \rho\dot{\theta}^2 - \rho\cos^2\theta\dot{\varphi}^2, \qquad a^{\theta} = \ddot{\theta} + \frac{2}{\rho}\dot{\rho}\dot{\varphi} + \sin\theta\cos\theta\dot{\varphi}^2,$$
$$a^{\varphi} = \ddot{\varphi} + \frac{2}{\rho}\dot{\rho}\dot{\varphi} - 2\tan\theta\dot{\theta}\dot{\varphi}.$$

Now, using (A.11), the *Newton's fundamental equation of motion, that is the basis of all science,* $\mathbf{F} = m\,\mathbf{a}$, gets the following tensorial form

$$F^i = ma^i = m\dot{\dot{v}}^i = m(v^i_{;k}\dot{x}^k) \equiv m(\dot{v}^i + \Gamma^i_{jk}v^j v^k) = m(\ddot{x}^i + \Gamma^i_{jk}\dot{x}^j\dot{x}^k), \quad \text{(A.12)}$$

which defines Newtonian force as a contravariant vector.

However, modern Hamiltonian dynamics reminds us that: (i) Newton's own force definition was not really $\mathbf{F} = m\,\mathbf{a}$, but rather $\mathbf{F} = \dot{\mathbf{p}}$, where $\mathbf{p}$ is the system's momentum, and (ii) the momentum $\mathbf{p}$ is not really a vector, but rather a dual quantity, a differential one–form[3] (see Chapter 2 for details). Consequently, the force, as its time derivative, is also a one–form (see Figure A.1). This new force definition includes the precise definition of the mass distribution within the system, by means of its Riemannian metric tensor g_{ij}. Thus, (A.12) has to be modified as

$$F_i = mg_{ij}a^j \equiv mg_{ij}(v^j_{;k}\dot{x}^k) \equiv mg_{ij}(\dot{v}^j + \Gamma^j_{ik}v^i v^k) = mg_{ij}(\ddot{x}^j + \Gamma^j_{ik}\dot{x}^i\dot{x}^k), \quad \text{(A.13)}$$

where the quantity mg_{ij} is called the *material metric tensor*, or *inertia matrix*. Equation (A.13) generalizes the notion of the Newtonian force $\mathbf{F}$, from Euclidean space $\mathbb{R}^n$ to the Riemannian manifold M (see Chapter 2).

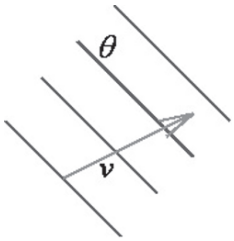

Fig. A.1. A one–form θ (which is a family of parallel (hyper)surfaces, the so–called *Grassmann planes*) pierced by the vector v to give a scalar product $\theta(v) \equiv < \theta, v > =$ 2.6 (see [MTW73] for technical details).

Application to Fluid Mechanics: Continuity Equation

The most important equation in continuum mechanics, in particular in fluid mechanics, is the celebrated *equation of continuity*,

[3] For example, in Dirac's $< bra|ket >$ formalism, *ket*s are vectors, while *bra*s are one–forms; in matrix notation, columns are vectors, while rows are one–forms.

$$\partial_t \rho + \mathrm{div}(\rho \dot{\mathbf{x}}) = 0. \tag{A.14}$$

Here we derive the continuity equation (A.14), using the absolute time derivative and starting from the *mass conservation principle*,

$$\overline{\dot{dm}} = 0, \tag{A.15}$$

where dm denotes an infinitesimal mass of a fluid (continuum) particle. If we introduce the fluid density $\rho = dm/dv$, where v is the infinitesimal volume of a fluid particle, then the mass conservation principle (A.15) can be written as

$$\overline{\dot{\rho dv}} = 0,$$

which is the absolute derivative of a product, and therefore expands into

$$\dot{\rho} dv + \rho \overline{\dot{dv}} = 0. \tag{A.16}$$

Now, as the fluid density is a function of both time t and spatial coordinates x^k, i.e., a scalar–field, $\rho = \rho(x^k, t)$, its total time derivative $\dot{\rho}$, figuring in (A.16), is defined by

$$\dot{\rho} = \partial_t \rho + \partial_{x^k} \rho \, \partial_t x^k \equiv \partial_t \rho + \rho_{;k} \dot{x}^k \equiv \partial_t \rho + \mathrm{grad}(\rho) \cdot \dot{\mathbf{x}}. \tag{A.17}$$

Regarding $\overline{\dot{dv}}$, the other term figuring in (A.16), we start by expanding an elementary volume dv along the sides $\{dx^i_{(p)}, dx^j_{(q)}, dx^k_{(r)}\}$ of an elementary parallelepiped,

$$dv = \frac{1}{3!} \delta^{pqr}_{ijk} dx^i_{(p)} dx^j_{(q)} dx^k_{(r)}, \qquad (i, j, k, p, q, r = 1, 2, 3)$$

so that its absolute derivative becomes

$$\begin{aligned} \overline{\dot{dv}} &= \frac{1}{2!} \delta^{pqr}_{ijk} \overline{\dot{dx^i}}_{(p)} dx^j_{(q)} dx^k_{(r)} \\ &= \frac{1}{2!} \dot{x}^i_{;l} \delta^{pqr}_{ijk} dx^l_{(p)} dx^j_{(q)} dx^k_{(r)} \qquad (\text{using } \overline{\dot{dx^i}}_{(p)} = \dot{x}^i_{;l} dx^l_{(p)}), \end{aligned}$$

which finally simplifies into

$$\overline{\dot{dv}} = \dot{x}^k_{;k} dv \equiv \mathrm{div}(\dot{\mathbf{x}})\, dv. \tag{A.18}$$

Substituting (A.17) and (A.18) into (A.16) gives

$$\overline{\dot{\rho dv}} \equiv \left(\partial_t \rho + \rho_{;k} \dot{x}^k\right) dv + \rho \dot{x}^k_{;k} dv = 0. \tag{A.19}$$

As we are dealing with arbitrary fluid particles, $dv \neq 0$, so from (A.19) follows

$$\partial_t \rho + \rho_{;k} \dot{x}^k + \rho \dot{x}^k_{;k} \equiv \partial_t \rho + (\rho \dot{x}^k)_{;k} = 0. \tag{A.20}$$

Equation (A.20) is the covariant form of the continuity equation, which in standard vector notation becomes (A.14).

In classical biomechanics, the continuity equation (A.14) forms the basis of *hemodynamics*, or blood flow dynamics.

A.1.4 The Covariant Force Law in Human–Like Biomechanics

Objective of this final tensor subsection is to generalize the fundamental Newtonian 3D equation, $\mathbf{F} = m\mathbf{a}$, for the generic biomechanical system, consisting of a number of flexibly–coupled rigid segments (see Figures 2.2–2.3 above), and thus to formulate the fundamental biomechanical law: *the covariant force law*.

To be able to apply the covariant formalism, we need to start with the suitable coordinate transformation (A.2), in this case as a relation between the 6 *external* $SE(3)$ rigid–body coordinates, $y^e = y^e(t)$ $(e = 1, ..., 6)$, and $2n$ *internal joint coordinates*, $x^i = x^i(t)$ $(i = 1, ..., 2n)$ (n angles, forming the constrained $n-$torus T^n, plus n very restricted translational coordinates, forming the hypercube $I^n \subset \mathbb{R}^n$). Once we have these two sets of coordinates, external–y^e and internal–x^i, we can perform the general functional transformation (A.2) between them,

$$x^i = x^i(y^e). \tag{A.21}$$

Now, although the coordinate transformation (A.21) is nonlinear and even unknown at this stage, there is something known and simple about it: the corresponding transformation of differentials is *linear and homogenous*,

$$dx^i = \frac{\partial x^i}{\partial y^e} dy^e,$$

which implies the *linear and homogenous transformation of velocities,*

$$\dot{x}^i = \frac{\partial x^i}{\partial y^e} \dot{y}^e. \tag{A.22}$$

Our *internal velocity vector–field* is defined by the set of ODEs (A.22), at each *representative point* $x^i = x^i(t)$ of the biomechanical configuration manifold $M = T^n \times I^n$, as $v^i \equiv v^i(x^i, t) := \dot{x}^i(x^i, t)$.

Note that in general, a *vector–field* represents a field of vectors defined at every point x^i within some region U (e.g., movable segments/joints only) of the total configuration manifold M (consisting of all the segments/joints). Analytically, vector–field is defined as a set of autonomous ODEs (in our case, the set (A.22)). Its solution gives the *flow*, consisting of *integral curves* of the vector–field, such that all the vectors from the vector–field are tangent to integral curves at different representative points $x^i \in U$. In this way, through every representative point $x^i \in U$ passes both a curve from the flow and its tangent vector from the vector–field. Geometrically, vector–field is defined as a cross–section of the tangent bundle TM, the so–called *velocity phase–space*. Its *geometrical dual* is the *1–form–field*, which represents a field of one–forms (see Figure A.1), defined at the same representative points $x^i \in U$. Analytically, 1–form–field is defined as an *exterior differential system*, an algebraic dual to the autonomous set of ODEs. Geometrically, it is defined as a cross–section of the

cotangent bundle T^*M, the so–called *momentum phase–space*. Together, the vector–field and its corresponding 1–form–field define the *scalar potential field* (e.g., kinetic and/or potential energy) at the same movable region $U \subset M$. See Chapter 2 for technical details.

Now, we are on the half–way to covariant biomechanics. We need to formulate the internal acceleration vector–field, $a^i \equiv a^i(x^i, \dot{x}^i, t)$, acting in all movable joints, and at the same time generalizing the Newtonian 3D acceleration vector **a**.

According to Newton, acceleration is a rate–of–change of velocity. But, from the previous subsections, we know that $a^i \neq \dot{v}^i$. However,

$$a^i := \dot{\tilde{v}}^i = \dot{v}^i + \Gamma^i_{jk} v^j v^k = \ddot{x}^i + \Gamma^i_{jk} \dot{x}^j \dot{x}^k. \tag{A.23}$$

Once we have the *internal acceleration vector–field* $a^i = a^i(x^i, \dot{x}^i, t)$, defined by the set of ODEs (A.23) (including *Levi–Civita connections* Γ^i_{jk} of the Riemannian configuration manifold M), we can finally define the *internal force 1–form field*, $F_i = F_i(x^i, \dot{x}^i, t)$, as a family of force one–forms, half of them rotational and half translational, acting in all movable joints,

$$F_i := m g_{ij} a^j = m g_{ij}(\dot{v}^j + \Gamma^j_{ik} v^i v^k) = m g_{ij}(\ddot{x}^j + \Gamma^j_{ik} \dot{x}^i \dot{x}^k), \tag{A.24}$$

where we have used the simplified *material metric tensor*, mg_{ij}, for the biomechanical system (considering, for simplicity, all segments to have equal mass m), defined by its *Riemannian kinetic energy form*

$$T = \frac{1}{2} m g_{ij} v^i v^j.$$

Equation $F_i = mg_{ij}a^j$, defined properly by (A.24) at every representative point x^i of the biomechanical configuration manifold M, formulates the sought for *covariant force law*, that generalizes the fundamental Newtonian equation, $\mathbf{F} = m\mathbf{a}$, for the generic biomechanical system. Its meaning is:

Force 1–form–field = Mass distribution × Acceleration vector–field

In other words, *the field (or, family) of force one–forms F_i, acting in all movable joints (with constrained rotations on T^n and very restricted translations on I^n), causes both rotational and translational accelerations of all body segments, within the mass distribution mg_{ij}*[4]*, along the flow–lines of the vector–field a^j.*

[4] More realistically, instead of the simplified metric mg_{ij} we have the *material metric tensor* G_{ij} (1.9), including all k segmental masses m_χ, as well as the corresponding moments and products of inertia,

$$G_{ij}(x, m) = \sum_{\chi=1}^{k} m_\chi \delta_{rs} \frac{\partial y^r}{\partial x^i} \frac{\partial y^s}{\partial x^j}, \qquad (r, s = 1, ..., 6;\ i, j = 1, ..., 2n),$$

as defined in Figures 2.2–2.3 above.

From the control theory perspective, a vector–field is a dynamical system, a set of differential equations (A.23) that has a set of force one–forms F_i as its inputs (see Chapter 5).

The purpose of Chapter 2 is to put this *core biomechanical law* into rigorous settings of smooth manifolds and their (co)tangent bundles.

A.1.5 The Essence of Hamiltonian Biomechanics

The *covariant force law*, $F_i = mg_{ij}a^j$, defined by (A.24) above, has the following Hamiltonian reformulation. We start with the *conservative Hamiltonian biomechanics* on the *cotangent bundle* T^*M of the system's *configuration manifold* M (see Figures 2.2–2.3 above), given by (see Chapter 3)

$$\dot{q}^\alpha = \partial_{p_\alpha} H(q^\alpha, p_\alpha), \qquad \dot{p}_\alpha = -\partial_{q^\alpha} H(q^\alpha, p_\alpha), \qquad (\alpha = 1, \dots, n).$$

The *forced Hamiltonian biomechanics* on T^*M is given by

$$\dot{q}^\alpha = \partial_{p_\alpha} H(q^\alpha, p_\alpha), \qquad \dot{p}_\alpha = F_\alpha(t, q^\alpha, p_\alpha) - \partial_{q^\alpha} H(q^\alpha, p_\alpha),$$

where F_α are muscular torques. The *generalized Hamiltonian biomechanics* (forced & dissipative) on T^*M is now given by

$$\begin{aligned} \dot{q}^\alpha &= \partial_{p_\alpha} H(q^\alpha, p_\alpha) - \partial_{p_\alpha} R(q^\alpha, p_\alpha), \\ \dot{p}_\alpha &= F_\alpha(t, q^\alpha, p_\alpha) - \partial_{q^\alpha} H(q^\alpha, p_\alpha) - \partial_{q^\alpha} R(q^\alpha, p_\alpha). \end{aligned} \tag{A.25}$$

The generalized Hamiltonian system (A.25) covers several types of classical dynamical systems (see Chapter 3):

(i) in case $F_\alpha = 0$, $R = 0$ and $H \neq 0$ – conservative Hamiltonian system;
(ii) in case $F_\alpha = 0$, $R \neq 0$ and $H \neq 0$ – dissipative Hamiltonian system;
(iii) in case $F_\alpha = 0$, $R \neq 0$ and $H = 0$ – bidirectional gradient system;
(iv) in case $F_\alpha \neq 0$, $R = 0$ and $H = 0$ – simple Newtonian system;
(v) in case $F_\alpha \neq 0$, $R = 0$ and $H \neq 0$ – generalized Newtonian system.

The *generalized Hamiltonian control system* on T^*M is obtained from (A.25) in the following way. First we introduce the *control Hamiltonian function*, $H_C : T^*M \times R \to R$. In the local coordinates $q^\alpha, p_\alpha \in U_p \subset T^*M$, the control Hamiltonian is given by

$$H_C(q, p, u) = H_0(q, p) - q^\alpha u_\alpha, \quad (\alpha = 1, \dots, n)$$

where $u_\alpha = u_\alpha(t, q, p)$ are *neural control inputs*, and the physical Hamiltonian $H_0(q, p)$ represents the system's *total energy function* $H_0 : T^*M \times \mathbb{R} \to \mathbb{R}$. The natural input–output control system is now defined as

$$\dot{q}^\alpha = \partial_{p_\alpha} H_C(q, p, u) + \partial_{p_\alpha} R(q, p), \qquad \dot{p}_\alpha = F_\alpha - \partial_{q^\alpha} H_C(q, p, u) + \partial_{q^\alpha} R(q, p),$$
$$y^\alpha = -\partial_{u_\alpha} H_C(q, p, u),$$

where y^α are *control outputs* (see Chapter 5).

A.2 Muscular System

A.2.1 Muscular Histology

Human skeletal and face muscles, accounting for more than 40% of the body weight in man, consist of bundles of elongated, cylindric cells called *muscle fibers*, 50 to 200 μ in diameter and often many centimeters long. Bundles of muscle fibers, each called *fasciculus*, are surrounded by a connective tissue covering, the *endomysium* (see, e.g., [Mou80, Mar98]).

A muscle consists of a number of fasciculi encased in a thick outer layer of connective tissue, the *perimysium*. At both ends of a muscle the connective tissue melds into a tendon by which the muscle is attached to the face or bony skeleton. In some muscles (*fusiform*), the muscle fibers run the whole length of muscle between the tendons, which form at opposite ends. In most muscles (*pennate*), one of the tendons penetrates through the center of the muscle; muscle fibers run at an angle to the axis of the whole muscle from the central tendon to the perimysium.

Like other cells, muscle cells are surrounded by a cell membrane, the *sarcolemma*. *Myofibrils*, the *contractile elements*, are numerous parallel, lengthwise threads 1 to 3 in diameter that fill most of the muscle fiber. The *cross striations*, seen in the skeletal and face muscles with electron microscope, are located in the myofibrils. Squeezed between the myofibrils and the sarcolemma is a small amount of cytoplasm, the *sarcoplasm*, in which are suspended multiple nuclei, numerous mitochondria, lysosomes, lipid droplets, glycogen granules, and other intracellular inclusions. The sarcoplasm contains glycogen, glycolytic enzymes, nucleotides, creatine phosphate, amino acids, and peptides.

Sarcoplasm also contains a well–developed endoplasmic reticulum, which in muscle is called *sarcoplasmic reticulum*. The sarcoplasmic reticulum forms an extensive hollow membranous system within the cytoplasm surrounding the myofibrils. Periodically, there are branching invaginations of the sarcolemma called *T tubules* or transverse tubules. The sarcoplasmic reticulum bulges out on either side of the T tubules to form large *lateral cisternae*. The T tubule and two sets of lateral cisternae constitute a *triad*. The triads play an important role in muscle excitation-contraction coupling (by release of Ca^{++} ions).

Two types of muscle fibers are found in human skeletal and face muscles: *red* and *white muscle fibers*, being histochemically and functionally distinctive. Many muscles are mixed, containing both types of fibers, which can be distinguished by various histochemical stains. In addition to muscle cells and fibroblasts in the connective tissue, a whole muscle contains fat cells and histiocytes.

Each muscle fiber contains numerous contractile elements - *myofibrils* ($1-3\,\mu$ in diameter) which are biological machines that utilize chemical energy from metabolism of food in the form of *adenosine triphosphate*, ATP hydrolysis to produce mechanical work. An understanding of contractility and muscle function requires, thus, both histo–mechanical and bio–energetic insight.

Contractile machinery unit of the myofibril, *sarcomere* ($1.5-3.5\,\mu$ long; on electron microscope it is seen as bounded by two Z lines, with H zone in the middle of the A band) is constituted of a great number of longitudinal protein filaments of two kinds: thick, *myosin* filaments (about $120\,\dot{A}$ in diameter and about $1.8\,\mu$ long; they are located in the center of the sarcomere arranged in a hexagonal array about $450\,\dot{A}$ apart) and thin, *actin* filaments (about $80\,\dot{A}$ in diameter and about $1.0\,\mu$ long; they are anchored into the transverse filaments forming the Z line) (see Figure A.2). Each myosin filament is surrounded by six actin filaments. Each myosin filament has two heads and two projections from opposite sides at about $143\,\dot{A}$ intervals along its length.

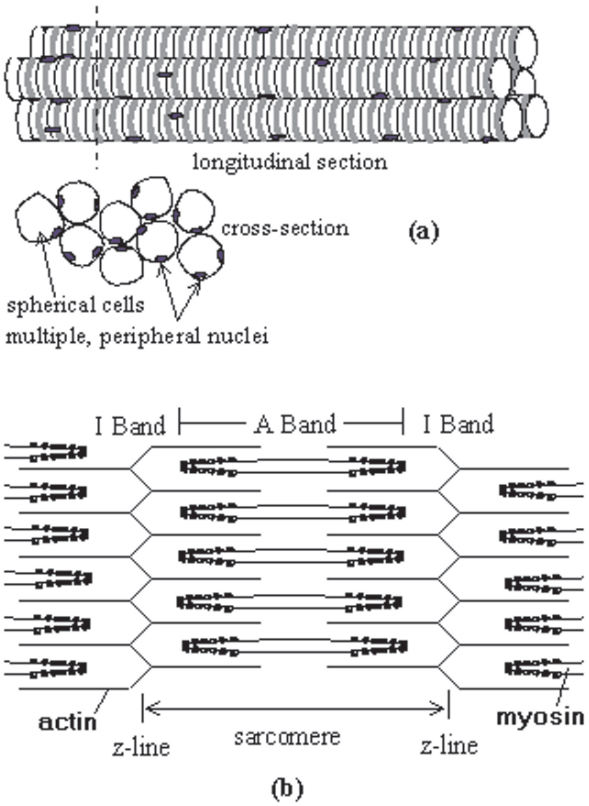

Fig. A.2. Cellular structure of the voluntary (skeletal) human muscle: (a) Muscular fibers with their cross–sections; (b) Sarcomere with overlapping myofilaments.

A.2.2 Classical Theories of Muscular Contraction

Huxley's Sliding Filament Theory

Essential for the contraction process are *cross bridges* (see Figure A.2). They extend from myosin filaments to touch one of the adjacent actin filaments. Each thin filament receives cross bridges from the three adjacent thick filaments. During shortening the two sets of interdigitating filaments slide with respect to each other, cross and finally overlap each other. This process of muscle shortening involving progressive interdigitation of the two sets of protein filaments represents the *sliding filament mechanism*, discovered and mathematically formulated as a *microscopic theory of muscular contraction* in 1954–57 by A.F. Huxley [HN54, Hux57].

According to Huxley, the myosin heads and cross bridges are elastic elements with a mechanism for attaching themselves transiently to specific sites on the thin filaments. The following cyclic events take place during muscular contraction:

1. The cross bridges extend from myosin filaments and attach themselves to specific sites on actin filaments. The probability that attachment will occur is $f(x)$, where x is the instantaneous distance between the equilibrium position (0) and the maximum distance for attachment h along the myofibrillar axis.
2. The cross bridges detach with probability $g(x)$.

If we let N equal the density of cross bridges and n the fraction of cross bridges that are attached, then nN equals the density of attached cross bridges. Huxley's rate equation for cross–bridge attachment–detachment, i.e. the *sliding filament model* of muscular contraction is now given by:

$$\dot{n} = f(x)[1 - n(x,t)] - g(x)n(x,t) = f(x) - [f(x) + g(x)]n(x,t). \quad \text{(A.26)}$$

Huxley's model (A.26) leads to expressions for the force developed by the cross bridges. For an *isometric steady–state contraction* the *contraction tension* or *contraction force* is given by:

$$F_0 = 0.5\, N\, h^2 \frac{kf}{f+g}, \quad \text{(A.27)}$$

where $k = k(x)$ is the stiffness of the cross–bridge spring. For *isotonic* steady states it recovers the classical *Hill's force–velocity* relation (A.28). The *static force* expression says that the force (or tension) generated in the muscle is the function of the interfilamentar overlap, and its maximum is about the middle of the shortening, where the acto–myosin overlap is maximal. This is the so–called *parabolic length–tension curve* of muscular contraction.

Hill's Force–Velocity Muscular Dynamics

The *dynamic force–velocity relation* of muscular contraction is firstly discovered in 1938, by A.V. Hill [Hil38], in his thermodynamic studies of muscular work, and put into the basis of *macroscopic muscle–load dynamics*. Hill's famous *hyperbolic force–velocity curve* has the equation:

$$(F + a)\, v \;=\; (F_0 + F)\, b, \tag{A.28}$$

and says that the muscle force is greatest in isometric conditions (without motion), while the velocity of shortening is maximal without external load; in other words, muscle is either 'strong' or 'fast', but no both. Constants a and b correspond respectively to the energy dissipated during the contraction and the velocity of the mechano–chemical processes.

Hill showed that energy change in muscle during contraction can be described by the following *thermodynamic relation*:

$$U \;=\; A \;+\; W \;+\; M, \tag{A.29}$$

where U is the total energy change associated with contraction, A is the *activation heat* (i.e., the heat production associated with the activation of the contractile elements), W is the mechanical work performed by the muscle by lifting a load, $\alpha \Delta x$ is the *shortening heat*, and M is the *maintenance heat* of contraction.

The activation heat begins and is almost completely liberated before any tension is developed, i.e. it is predominantly connected with the excitation–contraction coupling process, and corresponds in time to the *latency relaxation* of muscle. It is associated with the internal work required to transform the contractile elements from the resting to the active state. Part of the activation heat probably is associated with a change in the elastic properties of muscle, but about two thirds of it is associated with the release of Ca^{++} ions from the triads, its binding by troponin and the subsequent rearrangement of the thin filament proteins. The activation heat is greatest for the first twitch after a period rest and becomes smaller with succeeding twitches.

The maintenance heat begins at about the time tension begins and can be divided into two parts: the labile maintenance heat and the stable maintenance heat. For isometric contractions at shorter than rest length, both the labile and the stable heats diminish. For stretched muscle, the labile heat is approximately constant, whereas the stable heat diminishes with stretching and is roughly proportional to the degree of interfilamentar overlap. The stable heat has quite different values in functionally different muscles; it is law when the muscle maintains tension efficiently and vice versa.

The shortening heat is proportional mainly to the distance of shortening and does not depend greatly on the load, the speed of shortening, or the amount of work performed. Since mechanical work is $W \;=\; P \Delta x$, substituting this in the above thermodynamic relation (A.29) gives the *heat equation*:

$$U = A + (P + \alpha)\,\Delta x + M. \tag{A.30}$$

From the analogy of the term $(P + \alpha)$ in the heat equation (A.30) and the term $(P + a)$ in the force-velocity equation (A.28), Hill was able to show a rough equivalence between the coefficient of the shortening heat α and the force-velocity constant a. The shortening heat is greatest for the first twitch after a period of rest and is less for subsequent twitches.

Last, note should be made of *thermoelastic heat*. Generally speaking, resting muscle has rubberlike thermoelastic properties, whereas actively contracting muscle has springlike thermoelastic properties. During the development of tension the change in elastic properties is accompanied by an absorption of heat by the muscle. As tension falls during relaxation, an equivalent amount of heat is released by the muscle owing to its elastic properties. The various kinds of muscle heat must be corrected for the thermoelastic heat. However, for a complete cycle of contraction and relaxation, the net heat produced by thermoelastic mechanisms is zero.

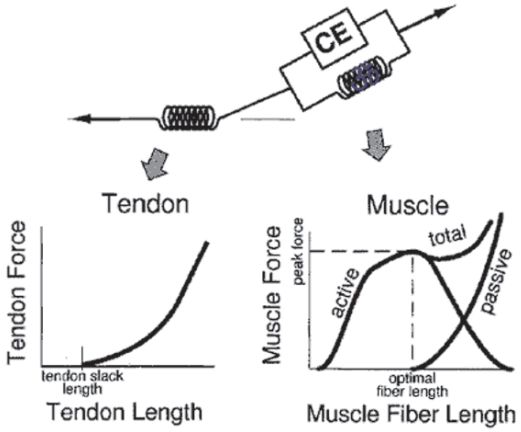

Fig. A.3. Hill's model of the skeletal muscle–tendon complex.

In the same seminal paper [Hil38], Hill also proposed a three–element rheological model of the skeletal muscle–tendon complex (see Figure A.3). In this model the length–tension property of muscle is represented by an active contractile element (CE) in parallel with a passive elastic element. Total isometric muscle force is assumed to be the sum of muscle force when it is inactive (passive) and when it is maximally excited (active). The muscle is in series with tendon, which is represented by a nonlinear spring. Pennation angle (α) is the angle between tendon and muscle fibers. Tendon slack length is the length of tendon at which force initially develops during tendon stretch. The model was scaled to represent each muscle by specifying the muscle's

peak force, optimal fiber length, tendon slack length, and pennation angle based on data collected in anatomical experiments.

Hill's muscle–tendon model has been widely applied in biomechanical musculo–skeletal modelling.

Hatze's Myocybernetics

Dynamics of human skeletal and face muscles is in the most sophisticated form described in the series of papers of Hatze (see [Hat78]). His muscle–control model involves excitation dynamics of neuro–muscular inputs (motor units) and contraction dynamics based on Huxley's sliding–filament theory of muscle contraction. In brief, Hatze's *myocybernetics* can be divided into *excitation dynamics* and *contraction dynamics*. The excitation dynamics of a single muscle fibre stimulated by trains of normalized nerve impulses $\alpha(t)$ is represented by the system

$$\ddot{\beta} + c_4\dot{\beta} + c_5\beta = c_6 V_N \alpha(t),\ \beta(0) = \dot{\beta}(0) = 0, \tag{A.31}$$

$$\ddot{\gamma} + (c_1\dot{\gamma} + c_2\gamma)/\rho^*(\xi) = c_3 V_T \beta(t),\ \gamma(0) = \dot{\gamma}(0) = 0,$$

$$\delta\dot{q} = d_1\{d_2[1 - k^2(\xi)][h(\dot{xi}) - 1/(1 - q_0)] - \delta q\}\,\delta q(t_s) = 0,$$

where $\rho^*(\xi)$ is *normalized Ca density function*, $k(\xi)$ is *filamentary–overlap function*, $h(\dot{xi})$ is *velocity–dependence function*;
$c_1, \ldots, c_6, d_1, d_2$, V_N, V_T, q_0 are defined constants; $V_T\beta(t)$ is *action potential* as appearing in the interior of the T–system of the fibre, while $\gamma(t)$ denotes the *free Ca–ion concentration* in the interfilamentary space; the variable δq expresses the *stretch potentiation* induced by an elongation of the tetanized fibre.

The variable ξ designates the *normalized length of the contractile element* of the fibre, and is defined by the contraction dynamics,

$$\dot{\xi} = a_1[1/a_2 \arcsin h\, a_3 \ln(\frac{q^* k(\xi)}{b_2[f^{SE}/\bar{f} + b_1 k_1(\xi)]} - a_4)] - \frac{1}{2},$$

$$\xi(0) = \xi_0, \tag{A.32}$$

where $a_1, \ldots, a_4, b_1, b_2$ are defined constants, $f^{SE}/\bar{f}$ is the *normalized force across the series elastic element*, $b_1 k_1(\xi)$ is the *passive sarcomere tension*, and q^* is the *active state*.

Hodgkin–Huxley Theory of Neural Action Potential

The celebrated *Hodgkin–Huxley HH–neuron model* is described by the nonlinear coupled differential equations for the four variables, V for the membrane potential, and m, h and n for the gating variables of Na and K channels, and it is given by [HH52, Hod64]

$$
\begin{aligned}
C\dot{V} &= -g_{\mathrm{Na}} m^3 h (V - V_{\mathrm{Na}}) - g_{\mathrm{K}} n^4 (V - V_{\mathrm{K}}) - g_{\mathrm{L}} (V - V_{\mathrm{L}}) + I_j^{\mathrm{ext}}, \\
\dot{m} &= -(a_m + b_m)\, m + a_m, \qquad \dot{h} = -(a_h + b_h)\, h + a_h, \qquad \text{(A.33)} \\
\dot{n} &= -(a_n + b_n)\, n + a_n, \qquad \text{where} \\
a_m &= 0.1\,(V + 40)/[1 - \mathrm{e}^{-(V+40)/10}], \qquad b_m = 4\,\mathrm{e}^{-(V+65)/18}, \\
a_n &= 0.01\,(V + 55)/[1 - \mathrm{e}^{-(V+55)/10}], \qquad b_n = 0.125\,\mathrm{e}^{-(V+65)/80}, \\
a_n &= 0.07\,\mathrm{e}^{-(V+65)/20}, \qquad b_n = 1/[1 + \mathrm{e}^{-(V+35)/10}].
\end{aligned}
$$

Here the reversal potentials of Na, K channels and leakage are $V_{\mathrm{Na}} = 50$ mV, $V_{\mathrm{K}} = -77$ mV and $V_{\mathrm{L}} = -54.5$ mV; the maximum values of corresponding conductivities are $g_{\mathrm{Na}} = 120$ mS/cm^2, $g_{\mathrm{K}} = 36$ mS/cm^2 and $g_{\mathrm{L}} = 0.3$ mS/cm^2; the capacity of the membrane is $C = 1\ \mu$F/cm^2. The external, input current is given by

$$
I_j^{\mathrm{ext}} = g_{syn}(V_a - V_c) \sum_n \alpha(t - t_{in}), \qquad \text{(A.34)}
$$

which is induced by the pre–synaptic spike–train input applied to the neuron i, given by

$$
U_{\mathrm{i}}(t) = V_{\mathrm{a}} \sum_n \delta(t - t_{\mathrm{in}}).
$$

In equation (A.34), t_{in} is the nth firing time of the spike–train inputs, g_{syn} and V_c denote the conductance and the reversal potential, respectively, of the synapse, τ_{s} is the time constant relevant to the synapse conduction, and $\alpha(t)$ is the alpha function given by

$$
\alpha(t) = (t/\tau_{\mathrm{s}})\ \mathrm{e}^{-t/\tau_{\mathrm{s}}} \Theta(t).
$$

where $\Theta(t)$ is the Heaviside function. The HH model was originally proposed to account for the property of squid giant axons [HH52, Hod64] and it has been generalized with modifications of ion conductances. The HH–type models have been widely adopted for a study on activities of *transducer neurons* such as motor and thalamus relay neurons, which transform the amplitude–modulated input to spike–train outputs.

Muscular Action Potential

Hodgkin–Huxley theory of neural action potential was adapted by Noble [Nob62] as a model of muscular action potential. Noble model has the same form as the HH–neuron model (A.33), with changed the values of constants, so that the whole signal is about 10 times slower. Noble's model was later modified by Hatze's muscular excitation dynamics (A.31) and complemented by his contraction dynamics (A.32).

Now, to simplify Hatze's myiocybernetics, and yet to retain all the necessary excitation–contraction dynamics, as well as to establish the neuro–muscular inter–connection, we propose herein approach of recurrent diffusion

physics. The EFS–response mapping $\mathcal{F}$ of a skeletal or face muscle, i.e., the response of the muscle system $\mathcal{M}$ to the efferent functional stimulation from the neural network system $\mathcal{N}$ – can be stated in the form of the *force generator* time behavior, $\mathcal{F} : \mathbb{R} \to Hom_t(\mathcal{N}, \mathcal{M})$, where: t denotes stimulation time, $\mathcal{N}$ and $\mathcal{M}$ correspond to the left $\mathbb{R}$–moduli of neural and muscular systems. The mapping $\mathcal{F}$ can be considered as an effect of a *fifth–order transmission cascade* $(\mathcal{F}_1 \mapsto \mathcal{F}_2 \mapsto \mathcal{F}_3 \mapsto \mathcal{F}_4 \mapsto \mathcal{F}_5)$, where $\mathcal{F}_i\,(i = 1, \ldots, 5)$ represent *neural action potential, synaptic potential, muscular action potential, excitation–contraction coupling and muscle tension generating*, respectively (see [Iva91]).

According to [Nob62, Hak93, Hak02], all transmission components of the system $(\mathcal{F}_1 \mapsto \mathcal{F}_2 \mapsto \mathcal{F}_3 \mapsto \mathcal{F}_4 \mapsto \mathcal{F}_5)$, where $\mathcal{F}_i\,(i = 1, \ldots, 5)$ can be considered as being some kind of *diffusion processes*, forming the fifth–order transmission *flux* cascade.

Mapping $\mathcal{F}$ (for all included motor units in the particular muscle contraction) can be described by fifth order *recurrent, distributed parameter* diffusion system [Iva91]

$$\begin{aligned} C_k \frac{\partial V_k}{\partial t} &= \frac{1}{R_k} \frac{\partial^2 V_{k-1}}{\partial z^2} - J_k(V_k), \qquad \text{with boundary condition at } z = 0, \\ V_k(0,t) &= V_0 \sin(2\pi f t) = S(t), \qquad (k = 1, \ldots, 5). \end{aligned}$$

The single element $\mathcal{F}_4$, $(k = 1, \ldots, 5)$ behavior is now given by

$$\begin{aligned} V_k(z,t) &= V_0 \exp(-z_k/m) \sin(2\pi f(t - z_k/n)), \\ m &= \frac{1}{R_k C_k f}, \qquad n = \frac{4\pi f}{R_k C_k}. \end{aligned}$$

For muscle–mechanical purpose, the presented distributed map $\mathcal{F}$ can be first mathematically approximated with the corresponding lumped parameter $R_k C_k$ electric circuit (where the second circuit represents the *Eccles model of synaptic activation* (see [Ecc64, EIS67]) and the last one corresponds to the low-pass filter representing the contraction process itself), at $x = tendon$

$$\begin{aligned} \dot{z}_k &= \frac{1}{T_k}(b_k z_{k-1} - z_k), \qquad (k = 1, \ldots, 5), \\ z_k(0) &= 0, \qquad z_0 = S(t), \qquad z_5 = F(t), \end{aligned}$$

where $T_k = R_k C_k$ are time characteristics of the circuits in cascade, and b_k are corresponding input gains (conversion factors).

The single muscle behavior in the lumped approximation form is given by the recurrent sum of its transient and weighting terms (with time legs τ_k)

$$z_k(t) = b_k z_{k-1}(1 - \exp(-t/T_k)) + z_k \exp(-(t - \tau_k)/T_k).$$

The presented distributed mapping $\mathcal{F}$ can be further physically approximated with a second order forced–dumped linear oscillator in a Cauchy form

$$T\ddot{z} + 2aT\dot{z} + cz = bS, \qquad z(0) = \dot{z}(0) = 0,$$

where a (having dimension of force) corresponds to energy dissipated during the contraction, b (having dimension of velocity) is the phosphagenic energy transducing rate, while c corresponds to the second derivative of the stress–strain curve of the series viscoelastic element [Wil56] of the muscular actuator (assumed in exponential three–parameter form).

The complete efferent face and body neuro–muscular system $(\mathcal{N}, \mathcal{M})$ is now given by the set of equations

$$\dot{x}^i = -D^i_j x^i + T^i_j g^i(x^i) + S^i, \qquad (i, j = 1, \ldots, n),$$
$$C_k \frac{\partial V_k}{\partial t} = \frac{1}{R_k} \frac{\partial^2 V_{k-1}}{\partial z^2} - J_k(V_k), \qquad (k = 1, \ldots, 5);$$

or, its discrete form

$$\dot{x}^i = (b - y^i)x^i, \qquad \dot{y}^i = -vy^i + g(x^i + T^i_j y^j), \tag{A.35}$$
$$\dot{z}_k = \frac{1}{T_k}(b_k z_{k-1} - z_k), \qquad (k = 1, \ldots, 5), \tag{A.36}$$
$$z_k(0) = 0, \qquad z_0 = S(t), \qquad z_5 = F(t). \tag{A.37}$$

Equations (A.35,A.37) constitute a $3n$–dimensional phase–space (for $n = 5$ or $k = i$) being a *hiper–cube* $\equiv$ *neuro–muscular control space*. The feedback control $\mathcal{F}^{-1}$ of the mapping $\mathcal{F}$ is performed by muscular *autogenetic motor servo*.

Houk's Autogenetic Motor Servo

It is now well–known (see [Hou79, HBB96]) that voluntary contraction force $\mathcal{F}$ of a skeletal or face muscle system $\mathcal{M}$ is reflexly excited (positive reflex feedback $+\mathcal{F}^{-1}$ by responses of its *spindle receptors* to stretch and is reflexly inhibited (negative reflex feedback $-\mathcal{F}^{-1}$ by responses of its *Golgi tendon organs* to contraction. Stretch and unloading reflexes are mediated by combined actions of several autogenetic neural pathways.

James Houk's term '*autogenetic*' means that the stimulus excites receptors located in the same face or body 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 $+\mathcal{F}^{-1}$, and the Golgi tendon organs, sensitive to contractile force - negative force feedback $-\mathcal{F}^{-1}$.

The gain G of the length feedback $+\mathcal{F}^{-1}$ can be expressed as the *positional stiffness* (the ratio $G \approx S = dF/dx$ of the force $\mathcal{F}$–change to the length $x-$change) of the muscle system $\mathcal{M}$. 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 $\mathcal{M}$ in executing controlled changes in length $+\mathcal{F}^{-1}$.

The autogenetic circuits $(+\mathcal{F}^{-1})$ and $(-\mathcal{F}^{-1})$ 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.

A.2.3 The Equivalent Muscular Actuator

A single skeletal muscle, (e.g., the triceps brachii muscle, see Figure A.4), is attached at its *origin* to a large area of bone (the humerus in case of the triceps). At its other end, the *insertion*, it tapers into a glistening white *tendon* which, (in case of the triceps is attached to the ulna). As the triceps contracts, the insertion is pulled toward the origin and the arm is straightened or extended at the elbow. Thus the triceps is an *extensor*. Because skeletal muscle exerts force only when it contracts, a second muscle – a *flexor* – is needed to flex or bend the joint (e.g., the biceps brachii muscle is the flexor of the forearm). Together, they (the biceps and triceps) make up an antagonistic pair of muscles, which we will call forming the *equivalent muscular actuator*. Similar pairs, i.e., equivalent muscular actuators, working antagonistically across other joints, provide for almost all the movement of the skeleton. The equivalent muscular actuator has the role of 'driver' in biodynamics. It generates the equivalent muscular torque, which is the *primary cause* of human–like motion [Iva91, IS01].

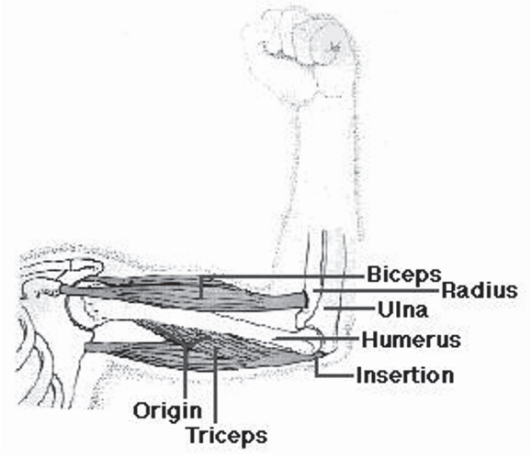

Fig. A.4. An antagonistic pair of human skeletal muscles, one flexor and the other extensor (in the case of the forearm, biceps brachii and triceps brachii, respectively), forming the *equivalent muscular actuator* – the primary cause of the human–like motion.

A.2.4 Biochemistry of Muscular Contraction

The immediate energy source for contraction in human muscles is adenosine triphosphate ATP (see [Mou80, Mar98, IS00]). Muscle contains about $2\,\mu mole$ ATP/gram wet weight. The myosin head is the only site of the major ATP hydrolysis in active muscle. At the concentrations of ATP, ADP (*adenosine diphosphate*), and P_i (*inorganic phosphate*) present in the sarcoplasm, ATP hydrolysis yields about $11.5\,kcal/mole$. About $0.3\,\mu mole$ of ATP/gram muscle is hydrolyzed by a single muscle twitch. The ATP hydrolysis overall scheme:

$$ATP \rightarrow ADP + P_i + \text{Energy} \tag{A.38}$$

represents actually the complex six–step chain–reaction $\{k_i\}$, $(i = 1, \ldots, 5)$, which can be summarized as follows:

1. Myosin reacts rapidly with ATP to form a complex; the myosin is from its resting (low–energy) form converted to an energy–rich form.
2. While complexed to the myosin, ATP is hydrolyzed to ADP and P_i. Reaction 2 is much more rapid than reaction 1. (This step is extremely temperature sensitive).
3. In reaction 3, while the ADP and P_i are still attached to the myosin, the latter is converted to a low-energy form. This step is slow, rate limiting in the sequence of reactions, and insensitive to temperature changes.
4. Reactions 4 and 5 are rapid.

Therefore, the reaction sequence proceeds as follows:

$$\begin{aligned} M + ATP \leftrightarrow M^* + ATP \leftrightarrow M^* + ADP + P \leftrightarrow \\ M + ADP + P, \leftrightarrow M + ADP + P_i \leftrightarrow M + ADP, \end{aligned} \tag{A.39}$$

where M is miosin in low–energy form, M^* is miosin in energy–rich form, and symbol $\leftrightarrow$ actually represents a pair of reversible reactions $\{k_i,\, k_{i-1}\}$, $(i = 1, \ldots, 5)$.

Muscles also contain about $20\,\mu mole$ CP/gram (*creatine phosphate*). Creatine phosphate can phosphorylate ADP to form ATP in a reversible reaction catalyzed by the enzyme *creatine kinase*.

Muscle contains large amounts of creatine kinase; it amounts to more than 25 percent of the soluble cytoplasmic protein. As soon as ATP is hydrolase, the ADP formed is very rapidly rephosphorylated by CP and the ATP is regenerated. Thus CP forms a reservoir of energy–rich phosphate bonds to quickly replenish the sarcoplasmic ATP.

Ultimately, ATP is produced by *glycolysis* and respiration [Mou80]. In glycolysis (the so called *Embden–Meyerhoff pathway*), glucose is degraded to pyruvate, or to lactic acid in the absence of O_2, yielding 2 moles $ATP/mole$ glucose metabolized. Intracellular glycogen granules provide a very readily available source of glucose. Muscles normally contain 9 to 16 gm/kg glycogen or, for a well–fed man of average height and weight, the total glycogen

stores in muscle amount to $300 - 500\,gram$, with another $55 - 90\,gram$ in the liver. Glycogen breakdown in muscle begins immediately on stimulation, and the amount of muscle glycogen depleted is proportional to the mechanical work done. Glycogen is hydrolyzed by the enzyme phosphorylase to *glucose–1–phosphate*, which then enters the glycolytic pathway.

Red muscle fibers respond to a stimulus with a relatively slow twitch (maximum shortening velocity about 17 mm/sec) and therefore are also called *slow fibers*, whereas *white muscle fibers* react to a stimulus with a rapid twitch (maximum shortening velocity about $42mm/sec$) and therefore are also called fast fibers. Red muscle has a more extensive blood supply than white muscle. Red muscle fibers are able to sustain activity for long periods of time whereas white muscle fibers characteristically produce short bursts of great tension followed by the rapid onset of fatigue.

Whole red and white muscles differ in *ATPase* activity, and, indeed, the purified contractile protein myosin extracted from red and white muscle differs in *ATPase* activity, a finding associated with different myosin light chains. White muscle and white muscle actomyosin show the greater *ATPase* activity. The innervation of red and white muscle differs, and, indeed, whether a given muscle is red or white results from trophic influences of the motor nerve.

Slow muscle fibers are generally thinner and possess many sarcosomes (mitochondria) containing large amounts of respiratory enzymes, as well as copious quantities of the O_2-carrying protein *myoglobin* in the sarcoplasm and man lipid droplets. The numerous sarcosomes and high level of myoglobin give slow fibers their red color. Fast (or white) muscle fibers, on the other hand, are generally of larger diameter and contain large amounts of phosphorylase and glycolytic enzymes and large deposits of glycogen. Slow muscles derive energy predominantly from respiration, whereas in fast muscle fibers, glycolysis and lactate production are more prominent.

A.3 Path Integral Methods

In this section we review Feynman path integral methods, from both historical and modern perspective (see also subsection 1.1.3).

A.3.1 Historical Remarks

Extract from Feynman's Nobel Lecture

In his Nobel Lecture, December 11, 1965, Richard (Dick) Feynman said that he and his PhD supervisor, John Wheeler, had found the *action* $A = A[x; t_i, t_j]$, directly involving the *motions of the charges only*,[5]

[5] *Wheeler–Feynman Idea* [WF49] "The energy tensor can be regarded only as a provisional means of representing matter. In reality, *matter consists of electrically charged particles*."

$$A[x;t_i,t_j] = m_i \int (\dot{x}^i_\mu \dot{x}^i_\mu)^{\frac{1}{2}} dt_i + \frac{1}{2} e_i e_j \int\int \delta(I^2_{ij})\, \dot{x}^i_\mu(t_i) \dot{x}^j_\mu(t_j)\, dt_i dt_j$$
$$\text{with } (i \neq j) \tag{A.40}$$
$$I^2_{ij} = \left[x^i_\mu(t_i) - x^j_\mu(t_j)\right] \left[x^i_\mu(t_i) - x^j_\mu(t_j)\right],$$

where $x^i_\mu = x^i_\mu(t_i)$ is the four–vector *position* of the ith particle as a function of the proper time t_i, while $\dot{x}^i_\mu(t_i) = dx^i_\mu(t_i)/dt_i$ is the *velocity* four–vector.

The first term in the action $A[x;t_i,t_j]$ (A.40) is the integral of the proper time t_i, the *ordinary action of relativistic mechanics of free particles of mass* m_i (summation over μ). The second term in the action $A[x;t_i,t_j]$ (A.40) represents the *electrical interaction of the charges*. It is summed over each pair of charges (the factor $\frac{1}{2}$ is to count each pair once, the term $i = j$ is omitted to avoid self–action). The *interaction is a double integral over a delta function of the square of space–time interval* I^2 *between two points on the paths*. Thus, interaction occurs only when this interval vanishes, that is, *along light cones* (see [WF45, WF49]).

Feynman comments here: "The fact that the interaction is exactly one–half advanced and half–retarded meant that we could write such a principle of least action, whereas interaction via retarded waves alone cannot be written in such a way. So, all of classical electrodynamics was contained in this very simple form."

"...The problem is only to make a quantum theory, which has as its classical analog, this expression (A.40). Now, there is no unique way to make a quantum theory from classical mechanics, although all the textbooks make believe there is. What they would tell you to do, was find the momentum variables and replace them by $(\hbar/i)(\partial/\partial x)$, but I couldn't find a momentum variable, as there wasn't any."

"The character of quantum mechanics of the day was to write things in the famous *Hamiltonian way* (in the form of Schrödinger equation), which described how the wave function changes from instant to instant, and in terms of the Hamiltonian operator H. If the classical physics could be reduced to a Hamiltonian form, everything was all right. Now, least action does not imply a Hamiltonian form if the action is a function of anything more than positions and velocities at the same moment. If the action is of the form of the integral of the Lagrangian $L = L(\dot{x}, x)$, a function of the velocities and positions at the same time t,

$$S[x] = \int L(\dot{x}, x)\, dt, \tag{A.41}$$

then you can start with the Lagrangian L and then create a Hamiltonian H and work out the quantum mechanics, more or less uniquely. But the action $A[x;t_i,t_j]$ (A.40) involves the key variables, positions (and velocities), at two different times t_i and t_j and therefore, it was not obvious what to do to make the quantum–mechanical analogue..."

So, Feynman was looking for the action integral in quantum mechanics. He says: "...I simply turned to Professor Jehle and said, "Listen, do you know

any way of doing quantum mechanics, starting with action – where the action integral comes into the quantum mechanics?" "No", he said, "but Dirac has a paper in which the Lagrangian, at least, comes into quantum mechanics." What Dirac said was the following: There is in quantum mechanics a very important quantity which carries the wave function from one time to another, besides the differential equation but equivalent to it, a kind of a kernel, which we might call $K(x', x)$, which carries the wave function $\psi(x)$ known at time t, to the wave function $\psi(x')$ at time $t + \varepsilon$,

$$\psi(x', t + \varepsilon) = \int K(x', x)\, \psi(x, t)\, dx.$$

Dirac points out that this function K was analogous to the quantity in classical mechanics that you would calculate if you took the exponential of $[i\varepsilon$ multiplied by the Lagrangian $L(\dot{x}, x)]$, imagining that these two positions x, x' corresponded to t and $t + \varepsilon$. In other words,

$$K(x', x) \quad \text{is analogous to} \quad \mathrm{e}^{i\varepsilon L(\frac{x'-x}{\varepsilon}, x)/\hbar}.$$

So, Feynman continues: "What does he mean, they are analogous; what does that mean, *analogous*? What is the use of that?" Professor Jehle said, "You Americans! You always want to find a use for everything!" I said that I thought that Dirac must mean that they were *equal*. "No", he explained, "he doesn't mean they are equal." "Well", I said, "Let's see what happens if we make them equal."

"So, I simply put them equal, taking the simplest example where the Lagrangian is

$$L = \frac{1}{2} M \dot{x}^2 - V(x),$$

but soon found I had to put a constant of proportionality N in, suitably adjusted. When I substituted for K to get

$$\psi(x', t + \varepsilon) = \int N \exp\left[\frac{i\varepsilon}{\hbar} L(\frac{x' - x}{\varepsilon}, x)\right] \psi(x, t)\, dx \tag{A.42}$$

and just calculated things out by Taylor series expansion, *out came the Schrödinger equation*. So, I turned to Professor Jehle, not really understanding, and said, "Well, you see, Dirac meant that they were proportional." Professor Jehle's eyes were bugging out – he had taken out a little notebook and was rapidly copying it down from the blackboard, and said, "No, no, this is an important discovery. You Americans are always trying to find out how something can be used. That's a good way to discover things!" So, I thought I was finding out what Dirac meant, but, as a matter of fact, had made the discovery that what Dirac thought was analogous, was, in fact, equal. I had then, at least, the connection between the Lagrangian and quantum mechanics, but still with wave functions and infinitesimal times."

"It must have been a day or so later when I was lying in bed thinking about these things, that I imagined what would happen if I wanted to calculate the wave function at a finite interval later. I would put one of these factors $e^{i\varepsilon L}$ in here, and that would give me the wave functions the next moment, $t+\varepsilon$, and then I could substitute that back into (A.42) to get another factor of $e^{i\varepsilon L}$ and give me the wave function the next moment, $t+2\varepsilon$, and so on and so on. In that way I found myself thinking of a large number of integrals, one after the other in sequence. In the integrand was the product of the exponentials, which, of course, was the exponential of the sum of terms like εL. Now, L is the Lagrangian and ε is like the time interval dt, so that if you took a sum of such terms, that's exactly like an integral. That's like Riemann's formula for the integral $\int L dt$, you just take the value at each point and add them together. We are to take the limit as $\varepsilon \to 0$, of course. Therefore, the connection between the wave function of one instant and the wave function of another instant a finite time later could be obtained by an infinite number of integrals (because ε goes to zero, of course), of exponential where S is the action expression (A.41). At last, I had succeeded in representing quantum mechanics directly in terms of the action $S[x]$."

Fully satisfied, Feynman comments: "This led later on to the idea of the **transition amplitude** for a path: that for each possible way that the particle can go from one point to another in space–time, there's an amplitude. That amplitude is **e** to the power of [$i/\hbar$ times the action $S[x]$ for the path], i.e., $e^{iS[x]/\hbar}$. Amplitudes from various paths superpose by addition. This then is another, a third way, of describing quantum mechanics, which looks quite different from that of Schrödinger or Heisenberg, but which is equivalent to them."

"...Now immediately after making a few checks on this thing, what I wanted to do, of course, was to substitute the action $A[x;t_i,t_j]$ (A.40) for the other $S[x]$ (A.41). The first trouble was that I could not get the thing to work with the relativistic case of spin one–half. However, although I could deal with the matter only nonrelativistically, I could deal with the light or the photon interactions perfectly well by just putting the interaction terms of (A.40) into any action, replacing the mass terms by the non–relativistic $Ldt = \frac{1}{2}M\dot{x}^2 dt$,

$$A[x;t_i,t_j] = \frac{1}{2}\sum_i m_i \int (\dot{x}_\mu^i)^2 dt_i + \frac{1}{2}\sum_{i,j(i\neq j)} e_i e_j \int\int \delta(I_{ij}^2)\, \dot{x}_\mu^i(t_i)\dot{x}_\mu^j(t_j)\, dt_i dt_j.$$

When the action has a delay, as it now had, and involved more than one time, I had to lose the idea of a wave function. That is, I could no longer describe the program as: given the amplitude for all positions at a certain time to compute the amplitude at another time. However, that didn't cause very much trouble. It just meant developing a new idea. *Instead of wave functions we could talk about this: that if a source of a certain kind emits a particle, and a detector is there to receive it, we can give the amplitude that the source will emit and the*

detector receive, $\mathrm{e}^{iA[x;t_i,t_j]/\hbar}$. We do this without specifying the exact instant that the *source* emits or the exact instant that any *detector* receives, without trying to specify the state of anything at any particular time in between, but by just finding the *amplitude for the complete experiment*. And, then we could discuss how that amplitude would change if you had a scattering sample in between, as you rotated and changed angles, and so on, without really having any wave functions...It was also possible to discover what the old concepts of energy and momentum would mean with this generalized action. And, so I believed that I had a quantum theory of classical electrodynamics – or rather of this new classical electrodynamics described by the action $A[x;t_i,t_j]$ (A.40)..."

Configuration (Lagrangian) Path Integral

Dirac and Feynman first developed the lagrangian approach to functional integration. To review this approach, we start with the *time–dependent Schrödinger equation*

$$i\hbar\,\partial_t\psi(x,t) = -\partial_{x^2}\psi(x,t) + V(x)\,\psi(x,t)$$

appropriate to a particle of mass m moving in a potential $V(x)$, $x \in \mathbb{R}$. A solution to this equation can be written as an integral (see e.g., [Kla97, Kla00]),

$$\psi(x'',t'') = \int K(x'',t'';x',t')\,\psi(x',t')\,dx' \ ,$$

which represents the wave function $\psi(x'',t'')$ at time t'' as a linear superposition over the wave function $\psi(x',t')$ at the initial time t', $t' < t''$. The integral kernel $K(x'',t'';x',t')$ is known as the *propagator*, and according to Feynman [Fey48] it may be given by

$$K(x'',t'';x',t') = \mathcal{N}\int \mathcal{D}[x]\ \mathrm{e}^{(i/\hbar)\int[(m/2)\,\dot{x}^2(t)-V(x(t))]\,dt},$$

which is a formal expression symbolizing an integral over a suitable set of paths. This integral is supposed to run over all continuous paths $x(t)$, $t' \leq t \leq t''$, where $x(t'') = x''$ and $x(t') = x'$ are fixed end points for all paths. Note that the integrand involves the *classical Lagrangian* for the system.

To overcome the convergence problems, Feynman adopted a *lattice regularization* as a procedure to yield well–defined integrals which was then followed by a limit as the lattice spacing goes to zero called the continuum limit. With $\varepsilon > 0$ denoting the lattice spacing, the details regarding the lattice regularization procedure are given by

$$K(x'',t'';x',t') = \lim_{\varepsilon\to 0}\,(m/2\pi i\hbar\varepsilon)^{(N+1)/2}\int\cdots$$
$$\cdots\int \exp\{(i/\hbar)\sum_{l=0}^{N}\,[(m/2\varepsilon)(x_{l+1}-x_l)^2 - \varepsilon\,V(x_l)\,]\}\,\prod_{l=1}^{N}\,dx_l\ ,$$

where $x_{N+1} = x''$, $x_0 = x'$, and $\varepsilon \equiv (t''-t')/(N+1)$, $N \in \{1, 2, 3, \dots\}$. In this version, at least, we have an expression that has a reasonable chance of being well defined, provided, of course, that one interprets the conditionally convergent integrals involved in an appropriate manner. One common and fully acceptable interpretation adds a convergence factor to the exponent of the preceding integral in the form $-(\varepsilon^2/2\hbar)\sum_{l=1}^{N} x_l^2$, which is a term that formally makes no contribution to the final result in the continuum limit save for ensuring that the integrals involved are now rendered absolutely convergent.

Phase–Space (Hamiltonian) Path Integral

It is necessary to retrace history at this point to recall the introduction of the *phase–space path integral* by Feynman [Fey51]. In Appendix B to this article, Feynman introduced a formal expression for the configuration or $q-$space propagator given by (see e.g., [Kla97, Kla00])

$$K(q'', t''; q', t') = \mathcal{M} \int \mathcal{D}[p]\, \mathcal{D}[q]\ \exp\{(i/\hbar) \int [\, p\, \dot{q} - H(p, q)\,]\, dt\}.$$

In this equation one is instructed to integrate over all paths $q(t)$, $t' \le t \le t''$, with $q(t'') \equiv q''$ and $q(t') \equiv q'$ held fixed, as well as to integrate over all paths $p(t)$, $t' \le t \le t''$, without restriction.

It is widely appreciated that the phase space path integral is more generally applicable than the original, Lagrangian, version of the path integral. For instance, the original configuration space path integral is satisfactory for Lagrangians of the general form

$$L(x) = \frac{1}{2}\, m\dot{x}^2 + A(x)\, \dot{x} - V(x)\ ,$$

but it is unsuitable, for example, for the case of a relativistic particle with the Lagrangian

$$L(x) = -m\, qrt1 - \dot{x}^2$$

expressed in units where the speed of light is unity. For such a system – as well as many more general expressions – the phase space form of the path integral is to be preferred. In particular, for the relativistic free particle, the phase space path integral

$$\mathcal{M} \int \mathcal{D}[p]\, \mathcal{D}[q]\ \exp\{(i/\hbar) \int [\, p\, \dot{q} -\ qrtp^2 + m^2\,]\, dt\},$$

is readily evaluated and yields the correct propagator.

Feynman–Kac Formula

Through his own research, M. Kac was fully aware of *Wiener's theory of Brownian motion* and the *associated diffusion equation* that describes the

corresponding *distribution function*. Therefore, it is not surprising that he was well prepared to give a path integral expression in the sense of Feynman for an equation similar to the time–dependent Schrödinger equation save for a rotation of the time variable by $-\pi/2$ in the complex plane, namely, by the change $t \to -it$ (see e.g., [Kla97, Kla00]). In particular, Kac [Kac51] considered the equation

$$\partial_t \rho(x,t) = \partial_{x^2} \rho(x,t) - V(x)\ \rho(x,t). \tag{A.43}$$

This equation is analogous to Schrödinger equation but of course differs from it in certain details. Besides certain constants which are different, and the change $t \to -it$, the nature of the dependent variable function $\rho(x,t)$ is quite different from the normal quantum mechanical wave function. For one thing, if the function ρ is initially real it will remain real as time proceeds. Less obvious is the fact that if $\rho(x,t) \geq 0$ for all x at some time t, then the function will continue to be nonnegative for all time t. Thus we can interpret $\rho(x,t)$ more like a probability density; in fact in the special case that $V(x) = 0$, then $\rho(x,t)$ is the probability density for a Brownian particle which underlies the *Wiener measure*. In this regard, ν is called the diffusion constant.

The fundamental solution of (A.43) with $V(x) = 0$ is readily given as

$$W(x,T;y,0) = \frac{1}{qrt2\pi\nu T}\ \exp\left(-\frac{(x-y)^2}{2\nu T}\right),$$

which describes the solution to the diffusion equation subject to the initial condition

$$\lim_{T\to 0^+}\ W(x,T;y,0) = \delta(x-y)\ .$$

Moreover, it follows that the solution of the diffusion equation for a general initial condition is given by

$$\rho(x'',t'') = \int W(x'',t'';x',t')\ \rho(x',t')\,dx'\ .$$

Iteration of this equation N times, with $\epsilon = (t''-t')/(N+1)$, leads to the equation

$$\rho(x'',t'') = N' \int \cdots \int \mathrm{e}^{-(1/2\nu\epsilon)\sum_{l=0}^{N}(x_{l+1}-x_l)^2}\ \prod_{l=1}^{N} dx_l\ \rho(x',t')\,dx',$$

where $x_{N+1} \equiv x''$ and $x_0 \equiv x'$. This equation features the imaginary time propagator for a free particle of unit mass as given formally as

$$W(x'',t'';x',t') = \mathcal{N} \int \mathcal{D}[x]\ \mathrm{e}^{-(1/2\nu)\int \dot{x}^2\,dt},$$

where $\mathcal{N}$ denotes a formal normalization factor.

The similarity of this expression with the Feynman path integral [for $V(x) = 0$] is clear, but there is a profound difference between these equations. In the former (Feynman) case the underlying measure is only *finitely additive*, while in the latter (Wiener) case the continuum limit actually defines a genuine measure, i.e., a *countably additive measure* on paths, which is a version of the famous *Wiener measure*. In particular,

$$W(x'', t''; x', t') = \int d\mu_W^\nu(x),$$

where μ_W^ν denotes a measure on continuous paths $x(t)$, $t' \leq t \leq t''$, for which $x(t'') \equiv x''$ and $x(t') \equiv x'$. Such a measure is said to be a *pinned* Wiener measure, since it specifies its path values at two time points, i.e., at $t = t'$ and at $t = t'' > t'$.

We note that Brownian motion paths have the property that with probability one they are concentrated on continuous paths. However, it is also true that the time derivative of a Brownian path is almost nowhere defined, which means that, with probability one, $\dot{x}(t) = \pm\infty$ for all t.

When the potential $V(x) \neq 0$ the propagator associated with (A.43) is formally given by

$$W(x'', t''; x', t') = \mathcal{N} \int \mathcal{D}[x] \mathrm{e}^{-(1/2\nu)\int \dot{x}^2 \, dt - \int V(x) \, dt},$$

an expression which is well defined if $V(x) \geq c$, $-\infty < c < \infty$. A mathematically improved expression makes use of the Wiener measure and is given by

$$W(x'', t''; x', t') = \int \mathrm{e}^{-\int V(x(t)) \, dt} \, d\mu_W^\nu(x).$$

This is an elegant relation in that it represents a solution to the differential equation (A.43) in the form of an integral over Brownian motion paths suitably weighted by the potential V. Incidentally, since the propagator is evidently a strictly positive function, it follows that the solution of the differential equation (A.43) is nonnegative for all time t provided it is nonnegative for any particular time value.

Itô Formula

Itô [Ito60] proposed another version of a *continuous–time regularization* that resolved some of the troublesome issues. In essence, the proposal of Itô takes the form given by

$$\lim_{\nu \to \infty} \mathcal{N}_\nu \int \mathcal{D}[x] \; \exp\{(i/\hbar) \int [\frac{1}{2} m\dot{x}^2 - V(x)] \, dt\} \exp\{-(1/2\nu) \int [\ddot{x}^2 + \dot{x}^2] \, dt\}.$$

Note well the alternative form of the auxiliary factor introduced as a regulator. The additional term $\ddot{x}^2$, the square of the second derivative of x, acts to smooth

out the paths sufficiently well so that in the case of (21) both $x(t)$ and $\dot{x}(t)$ are continuous functions, leaving $\ddot{x}(t)$ as the term which does not exist. However, since only x and $\dot{x}$ appear in the rest of the integrand, the indicated path integral can be well defined; this is already a positive contribution all by itself (see e.g., [Kla97, Kla00]).

A.3.2 Standard Path Integral Quantization

Canonical versus Path Integral Quantization

Recall that in the usual, *canonical formulation* of quantum mechanics, the system's phase–space coordinates, q, and momenta, p, are replaced by the corresponding Hermitian operators in the Hilbert space, with real measurable eigenvalues, which obey *Heisenberg commutation relations*.

The *path integral quantization* is instead based directly on the notion of a propagator $K(q_f, t_f; q_i, t_i)$ which is defined such that (see [Ryd96, CL84, Gun03])

$$\psi(q_f, t_f) = \int K(q_f, t_f; q_i, t_i)\, \psi(q_i, t_i)\, dq_i, \tag{A.44}$$

i.e., the wave function $\psi(q_f, t_f)$ at final time t_f is given by a Huygens principle in terms of the wave function $\psi(q_i, t_i)$ at an initial time t_i, where we have to integrate over all the points q_i since all can, in principle, send out little wavelets that would influence the value of the wave function at q_f at the later time t_f. This equation is very general and is simply an expression of causality. We use the normal units with $\hbar = 1$.

According to the usual interpretation of quantum mechanics, $\psi(q_f, t_f)$ is the *probability amplitude* that the particle is at the point q_f and the time t_f, which means that $K(q_f, t_f; q_i, t_i)$ is the probability amplitude for a transition from q_i and t_i to q_f and t_f. The probability that the particle is observed at q_f at time t_f if it began at q_i at time t_i is

$$P(q_f, t_f; q_i, t_i) = \left| K(q_f, t_f; q_i, t_i) \right|^2 .$$

Let us now divide the time interval between t_i and t_f into two, with t as the intermediate time, and q the intermediate point in space. Repeated application of (A.44) gives

$$\psi(q_f, t_f) = \int\int K(q_f, t_f; q, t)\, dq\, K(q, t; q_i, t_i)\, \psi(q_i, t_i)\, dq_i,$$

from which it follows that

$$K(q_f, t_f; q_i, t_i) = \int dq\, K(q_f, t_f; q, t)\, K(q, t; q_i, t_i).$$

This equation says that the transition from (q_i, t_i) to (q_f, t_f) may be regarded as the result of the transition from (q_i, t_i) to all available intermediate points

q followed by a transition from (q,t) to (q_f,t_f). This notion of *all possible paths* is crucial in the path integral formulation of quantum mechanics.

Now, recall that the *state vector* $|\psi,t\rangle_S$ in the *Schrödinger picture* is related to that in the *Heisenberg picture* $|\psi\rangle_H$ by

$$|\psi,t\rangle_S = \mathrm{e}^{-iHt}\,|\psi\rangle_H\,,$$

or, equivalently,

$$|\psi\rangle_H = \mathrm{e}^{iHt}\,|\psi,t\rangle_S\,.$$

We also define the vector

$$|q,t\rangle_H = \mathrm{e}^{iHt}\,|q\rangle_S\,,$$

which is the Heisenberg version of the Schrödinger state $|q\rangle$. Then, we can equally well write

$$\psi(q,t) = \langle q,t\;|\psi\rangle_H\,. \tag{A.45}$$

By completeness of states we can now write

$$\langle q_f,t_f\;|\psi\rangle_H = \int \langle q_f,t_f\;|q_i,t_i\rangle_H\;\langle q_i,t_i\;|\psi\rangle_H\;dq_i,$$

which with the definition of (A.45) becomes

$$\psi(q_f,t_f) = \int \langle q_f,t_f\;|q_i,t_i\rangle_H\;\psi(q_i,t_i)\,dq_i.$$

Comparing with (A.44), we get

$$K(q_f,t_f;q_i,t_i) = \langle q_f,t_f\;|q_i,t_i\rangle_H\,.$$

Now, let us compute the *quantum–mechanics propagator*

$$\langle q',t'\;|q,t\rangle_H = \left\langle q'|\mathrm{e}^{-iH(t-t')}\;|q\right\rangle$$

using the path integral formalism that will incorporate the direct quantization of the coordinates, without Hilbert space and Hermitian operators.

The first step is to divide up the time interval into $n+1$ tiny pieces: $t_l = l\varepsilon + t$ with $t' = (n+1)\varepsilon + t$. Then, by completeness, we can write (dropping the Heisenberg picture index H from now on)

$$\begin{aligned}\langle q',t'\;|q,t\rangle = \int dq_1(t_1)... \int dq_n(t_n)\;\langle q',t'\;|q_n,t_n\rangle\;\times \\ \times\,\langle q_n,t_n\;|q_{n-1},t_{n-1}\rangle\,...\,\langle q_1,t_1\;|q,t\rangle\,.\end{aligned} \tag{A.46}$$

The integral $\int dq_1(t_1)...dq_n(t_n)$ is an *integral over all possible paths*, which are not trajectories in the normal sense, since there is no requirement of continuity, but rather *Markov chains*.

Now, for small ε we can write

$$\langle q', \varepsilon \, |q, 0\rangle = \left\langle q' | \mathrm{e}^{-i\varepsilon H(P,Q)} \, |q\right\rangle = \delta(q'-q) - i\varepsilon \, \langle q' | H(P,Q) \, |q\rangle \, ,$$

where $H(P,Q)$ is the Hamiltonian (e.g., $H(P,Q) = \frac{1}{2}P^2 + V(Q)$, where P, Q are the momentum and coordinate operators). Then we have (see [Ryd96, CL84, Gun03])

$$\langle q' | H(P,Q) \, |q\rangle = \int \frac{dp}{2\pi} \, \mathrm{e}^{ip(q'-q)} H\left(p, \frac{1}{2}(q'+q)\right).$$

Putting this into our earlier form we get

$$\langle q', \varepsilon \, |q, 0\rangle \simeq \int \frac{dp}{2\pi} \exp\left[i \left\{ p(q'-q) - \varepsilon H\left(p, \frac{1}{2}(q'+q)\right)\right\}\right],$$

where the 0th order in $\varepsilon \to \delta(q'-q)$ and the 1st order in $\varepsilon \to -i\varepsilon \, \langle q' | H(P,Q) \, |q\rangle$. If we now substitute many such forms into (A.46) we finally get

$$\langle q', t' \, |q, t\rangle = \lim_{n\to\infty} \int \prod_{i=1}^{n} dq_i \prod_{k=1}^{n+1} \frac{dp_k}{2\pi} \times \tag{A.47}$$

$$\times \exp\left\{ i \sum_{j=1}^{n+1} [p_j(q_j - q_{j-1})] - H\left(p_j, \frac{1}{2}(q_j + q_{j+1})\right)(t_j - t_{j-1})]\right\},$$

with $q_0 = q$ and $q_{n+1} = q'$. Roughly, the above formula says to *integrate over all possible momenta and coordinate values associated with a small interval*, weighted by something that is going to turn into the *exponential of the action* e^{iS} in the limit where $\varepsilon \to 0$. It should be stressed that the different q_i and p_k integrals are independent, which implies that p_k for one interval can be completely different from the $p_{k'}$ for some other interval (including the neighboring intervals). In principle, the integral (A.47) should be defined by *analytic continuation into the complex plane* of, for example, the p_k integrals.

Now, if we go to the differential limit where we call $t_j - t_{j-1} \equiv d\tau$ and write $\frac{(q_j - q_{j-1})}{(t_j - t_{j-1})} \equiv \dot{q}$, then the above formula takes the form

$$\langle q', t' \, |q, t\rangle = \int \mathcal{D}[p]\mathcal{D}[q] \exp\left\{ i \int_t^{t'} [p\dot{q} - H(p,q)] \, d\tau \right\},$$

where we have used the shorthand notation

$$\int \mathcal{D}[p]\mathcal{D}[q] \equiv \int \prod_{\tau} \frac{dq(\tau)dp(\tau)}{2\pi}.$$

Note that the above integration is an integration over the p and q values at every time τ. This is what we call a *functional integral*. We can think of a

given set of choices for all the $p(\tau)$ and $q(\tau)$ as defining a *path in the* $6D$ *phase–space*. The most important point of the above result is that we have obtained an expression for a *quantum–mechanical transition amplitude* in terms of an integral involving only pure complex numbers, without operators.

We can actually perform the above integral for Hamiltonians of the type $H = H(P, Q)$. We use square completion in the exponential for this, defining the integral in the complex p plane and continuing to the physical situation. In particular, we have

$$\int_{-\infty}^{\infty} \frac{dp}{2\pi} \exp\left\{ i\varepsilon(p\dot{q} - \frac{1}{2}p^2] \right\} = \frac{1}{\sqrt{2\pi i\varepsilon}} \exp\left[\frac{1}{2} i\varepsilon \dot{q}^2 \right],$$

(see [Ryd96, CL84, Gun03]) which, substituting into (A.47) gives

$$\langle q', t' \,|q, t\rangle = \lim_{n\to\infty} \int \prod_i \frac{dq_i}{\sqrt{2\pi i\varepsilon}} \exp\left\{ i\varepsilon \sum_{j=1}^{n+1} \left[\frac{1}{2} \left(\frac{q_j - q_{j-1}}{\varepsilon} \right)^2 - V\left(\frac{q_j + q_{j+1}}{2} \right) \right] \right\}.$$

This can be formally written as

$$\langle q', t' \,|q, t\rangle = \int \mathcal{D}[q]\, \mathrm{e}^{iS[q]},$$

where

$$\int \mathcal{D}[q] \equiv \int \prod_i \frac{dq_i}{\sqrt{2\pi i\varepsilon}},$$

while

$$S[q] = \int_t^{t'} L(q, \dot{q})\, d\tau$$

is the *standard action* with the *Lagrangian*

$$L = \frac{1}{2}\dot{q}^2 - V(q).$$

Generalization to many degrees of freedom is straightforward:

$$\langle q'_1 ... q'_N, t' | q_1 ... q_N, t\rangle = \int \mathcal{D}[p]\mathcal{D}[q] \exp\left\{ i \int_t^{t'} \left[\sum_{n=1}^{N} p_n \dot{q}_n - H(p_n, q_n) \right] d\tau \right\},$$

$$\text{with} \qquad \int \mathcal{D}[p]\mathcal{D}[q] = \int \prod_{n=1}^{N} \frac{dq_n dp_n}{2\pi}.$$

Here, $q_n(t) = q_n$ and $q_n(t') = q'_n$ for all $n = 1, ..., N$, and we are allowing for the full Hamiltonian of the system to depend upon all the N momenta and coordinates collectively.

Elementary Applications

(i) Consider first

$$\begin{aligned}&\langle q', t'|Q(t_0)|q, t\rangle \\ &= \int \prod dq_i(t_i)\, \langle q', t'|q_n, t_n\rangle \ldots \langle q_{i0}, t_{i0}|Q(t_0)|q_{i-1}, t_{i-1}\rangle \ldots \langle q_1, t_1|q, t\rangle\,,\end{aligned}$$

where we choose one of the time interval ends to coincide with t_0, i.e., $t_{i0} = t_0$. If we operate $Q(t_0)$ to the left, then it is replaced by its eigenvalue $q_{i0} = q(t_0)$. Aside from this one addition, everything else is evaluated just as before and we will obviously obtain

$$\langle q', t'|Q(t_0)|q, t\rangle = \int \mathcal{D}[p]\mathcal{D}[q]\, q(t_0) \exp\left\{ i \int_t^{t'} [p\dot{q} - H(p, q)] d\tau \right\}.$$

(ii) Next, suppose we want a path integral expression for $\langle q', t'|Q(t_1)Q(t_2)|q, t\rangle$ in the case where $t_1 > t_2$. For this, we have to insert as intermediate states $|q_{i1}, t_{i1}\rangle \langle q_{i1}, t_{i1}|$ with $t_{i1} = t_1$ and $|q_{i2}, t_{i2}\rangle \langle q_{i2}, t_{i2}|$ with $t_{i2} = t_2$ and since we have ordered the times at which we do the insertions we must have the first insertion to the left of the 2nd insertion when $t_1 > t_2$. Once these insertions are done, we evaluate $\langle q_{i1}, t_{i1}|\, Q(t_1) = \langle q_{i1}, t_{i1}|\, q(t_1)$ and $\langle q_{i2}, t_{i2}|\, Q(t_2) = \langle q_{i2}, t_{i2}|\, q(t_2)$ and then proceed as before and get

$$\langle q', t'|Q(t_1)Q(t_2)|q, t\rangle = \int \mathcal{D}[p]\mathcal{D}[q]\, q(t_1)\, q(t_2) \exp\left\{ i \int_t^{t'} [p\dot{q} - H(p, q)] d\tau \right\}.$$

Now, let us ask what the above integral is equal to if $t_2 > t_1$? It is obvious that what we get for the above integral is $\langle q', t'|Q(t_2)Q(t_1)|q, t\rangle$. Clearly, this generalizes to an arbitrary number of Q operators.

(iii) When we enter into quantum field theory, the Q's will be replaced by fields, since it is the fields that play the role of coordinates in the 2nd quantization conditions.

Sources

The *source* is represented by modifying the Lagrangian:

$$L \to L + J(t)q(t).$$

Let us define $|0, t\rangle^J$ as the ground state (vacuum) vector (in the moving frame, i.e. with the e^{iHt} included) in the presence of the source. The required *transition amplitude* is

$$Z[J] \propto \langle 0, +\infty|0, -\infty\rangle^J\,,$$

where the source $J = J(t)$ plays a role analogous to that of an electromagnetic current, which acts as a source of the electromagnetic field. In other words, we

can think of the scalar product $J_\mu A^\mu$, where J_μ is the current from a scalar (or Dirac) field acting as a source of the potential A^μ. In the same way, we can always define a current J that acts as the source for some arbitrary field ϕ. $Z[J]$ (otherwise denoted by $W[J]$) is a functional of the current J, defined as (see [Ryd96, CL84, Gun03])

$$Z[J] \propto \int \mathcal{D}[p]\mathcal{D}[q] \exp\left\{ i \int_t^{t'} [p(\tau)\dot{q}(\tau) - H(p,q) + J(\tau)q(\tau)]d\tau \right\},$$

with the *normalization condition* $Z[J=0]=1$. Here, the argument of the exponential depends upon the functions $q(\tau)$ and $p(\tau)$ and we then integrate over all possible forms of these two functions. So the exponential is a functional that maps a choice for these two functions into a number. For example, for a quadratically completable $H(p,q)$, the p integral can be performed as a q integral

$$Z[J] \propto \int \mathcal{D}[q] \exp\left\{ i \int_{-\infty}^{+\infty} \left(L + Jq + \frac{1}{2} i\varepsilon q^2 \right) d\tau \right\},$$

where the addittion to H was chosen in the form of a *convergence factor* $-\frac{1}{2}i\varepsilon q^2$.

Fields

Let us now treat the *abstract scalar field* $\phi(x)$ as a coordinate in the sense that we imagine dividing space up into many little cubes and the average value of the field $\phi(x)$ in that cube is treated as a coordinate for that little cube. Then, we go through the multi–coordinate analogue of the procedure we just considered above and take the continuum limit. The final result is

$$Z[J] \propto \int \mathcal{D}[\phi] \exp\left\{ i \int d^4x \left(\mathcal{L}\left(\phi(x)\right) + J(x)\phi(x) + \frac{1}{2} i\varepsilon \phi^2 \right) \right\},$$

where for $\mathcal{L}$ we would employ the Klein Gordon Lagrangian form. In the above, the dx_0 integral is the same as $d\tau$, while the $d^3\mathbf{x}$ integral is simply summing over the sub–Lagrangians of all the different little cubes of space and then taking the continuum limit. $\mathcal{L}$ is the Lagrangian density describing the Lagrangian for each little cube after taking the many–cube limit (see [Ryd96, CL84, Gun03]) for the full derivation).

We can now introduce *interactions,* $\mathcal{L}_I$. Assuming the simple form of the Hamiltonian, we have

$$Z[J] \propto \int \mathcal{D}[\phi] \exp\left\{ i \int d^4x \left(\mathcal{L}\left(\phi(x)\right) + \mathcal{L}_I\left(\phi(x)\right) + J(x)\phi(x) \right) \right\},$$

again using the normalization factor required for $Z[J=0]=1$.

For example of Klein Gordon theory, we would use

$$\mathcal{L} = \mathcal{L}_0 + \mathcal{L}_I, \qquad \mathcal{L}_0 \frac{1}{2}[\partial_\mu \phi \partial^\mu \phi - \mu^2 \phi^2], \qquad \mathcal{L}_I = \mathcal{L}_I(\phi),$$

where $\partial_\mu \equiv \partial_{x^\mu}$ and we can freely manipulate indices, as we are working in Euclidean space $\mathbb{R}^3$. In order to define the above $Z[J]$, we have to include a convergence factor $i\varepsilon\phi^2$,

$$\mathcal{L}_0 \to \frac{1}{2}[\partial_\mu \phi \partial^\mu \phi - \mu^2 \phi^2 + i\varepsilon\phi^2],$$

so that

$$Z[J] \propto \int \mathcal{D}[\phi] \exp\left\{ i \int d^4x \left(\frac{1}{2}[\partial_\mu \phi \partial^\mu \phi - \mu^2 \phi^2 + i\varepsilon\phi^2] + \mathcal{L}_I\left(\phi(x)\right) + J(x)\phi(x) \right) \right\}$$

is the appropriate *generating function* in the free field theory case.

Gauges

In the path integral approach to quantization of the *gauge theory*, we implement *gauge fixing* by restricting in some manner or other the path integral over gauge fields $\int \mathcal{D}[A_\mu]$. In other words we will write instead

$$Z[J] \propto \int \mathcal{D}[A_\mu]\, \delta\left(\text{some gauge fixing condition}\right) \exp\left\{ i \int d^4x \mathcal{L}\left(A_\mu\right) \right\}.$$

A common approach would be to start with the *gauge condition*

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} - \frac{1}{2} (\partial^\mu A_\mu)^2$$

where the electrodynamic field tensor is given by $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, and compute

$$Z[J] \propto \int \mathcal{D}[A_\mu] \exp\left\{ i \int d^4x \left[\mathcal{L}(A_\mu(x)) + J_\mu(x) A^\mu(x) \right] \right\}$$

as the *generating function* for the *vacuum expectation* values of *time ordered products* of the A_μ fields. Note that J_μ should be conserved ($\partial^\mu J_\mu = 0$) in order for the full expression $\mathcal{L}(A_\mu) + J_\mu A^\mu$ to be *gauge invariant* under the integral sign when $A_\mu \to A_\mu + \partial^\mu \Lambda$. For a proper approach, see [Ryd96, CL84, Gun03].

Geometries

In this subsection, following [SK98], we describe path integral quantization on Riemannian–symplectic manifolds. Let $\hat{q}^j$ be a set of Cartesian coordinate canonical operators satisfying the Heisenberg commutation relations $[\hat{q}^j, \hat{q}^k] = i\omega^{jk}$. Here $\omega^{jk} = -\omega^{kj}$ is the canonical symplectic structure (see section 2.6.2 above). We introduce the canonical coherent states as $|q\rangle \equiv \mathrm{e}^{iq^j \omega_{jk} \hat{q}^k} |0\rangle$, where

$\omega_{jn}\omega^{nk} = \delta_j^k$, and $|0\rangle$ is the ground state of a harmonic oscillator with unit angular frequency. Any state $|\psi\rangle$ is given as a function on phase space in this representation by $\langle q|\psi\rangle = \psi(q)$. A general operator $\hat{A}$ can be represented in the form $\hat{A} = \int dq\, a(q)|q\rangle\langle q|$, where $a(q)$ is the lower symbol of the operator and dq is a properly normalized form of the Liouville measure. The function $A(q,q') = \langle q|\hat{A}|q'\rangle$ is the *kernel of the operator*.

The main object of the path integral formalism is the integral kernel of the evolution operator

$$K_t(q,q') = \langle q|\mathrm{e}^{-it\hat{H}}|q'\rangle = \int\limits_{q(0)=q'}^{q(t)=q} \mathcal{D}[q]\ \mathrm{e}^{i\int_0^t d\tau\left(\frac{1}{2}q^j\omega_{jk}\dot{q}^k - h\right)}\ . \tag{A.48}$$

Here $\hat{H}$ is the Hamiltonian, and $h(q)$ its symbol. The measure formally implies a sum over all phase-space paths pinned at the initial and final points, and a Wiener measure regularization implies the following replacement

$$\mathcal{D}[q] \to \mathcal{D}[\mu_\nu(q)] = \mathcal{D}[q]\,\mathrm{e}^{-\frac{1}{2\nu}\int_0^t d\tau\,\dot{q}^2} = N_\nu(t)\,d\mu_W^\nu(q)\ . \tag{A.49}$$

The factor $N_\nu(t)$ equals $2\pi\mathrm{e}^{\nu t/2}$ for every degree of freedom, $d\mu_W^\nu(q)$ stands for the Wiener measure, and ν denotes the diffusion constant. We denote by $K_t^\nu(q,q')$ the integral kernel of the evolution operator for a finite ν. The Wiener measure determines a stochastic process on the *flat* phase space. The integral of the symplectic one-form $\int q\omega dq$ is a stochastic integral that is interpreted in the Stratonovich sense. Under general coordinate transformations $q = q(\bar{q})$, the Wiener measure describes *the same* stochastic process on *flat* space in the curvilinear coordinates $dq^2 = d\sigma(\bar{q})^2$, so that the value of the integral is not changed apart from a possible phase term. After the calculation of the integral, the evolution operator kernel is obtained by taking the limit $\nu \to \infty$. The existence of this limit, and also the covariance under general phase-space coordinate transformations, can be most easily proved through the *operator* formalism for the regularized kernel $K_t^\nu(q,q')$.

Note that the integral (A.48) with the Wiener measure inserted can be regarded as an ordinary Lagrangian path integral with a complex action, where the configuration space is the original phase space and the Hamiltonian $h(q)$ serves as a potential. Making use of this observation it is not hard to derive the corresponding Schrödinger–like equation

$$\partial_t K_t^\nu(q,q') = \left[\frac{\nu}{2}\left(\partial_{q^j} + \frac{i}{2}\omega_{jk}q^k\right)^2 - ih(q)\right] K_t^\nu(q,q')\ , \tag{A.50}$$

subject to the initial condition $K_{t=0}^\nu(q,q') = \delta(q-q'), 0 < \nu < \infty$. One can easily show that $\hat{K}_t^\nu \to \hat{K}_t$ as $\nu \to \infty$ for all $t > 0$. The covariance under general coordinate transformations follows from the covariance of the "kinetic" energy of the Schrödinger operator in (A.50): The Laplace operator is replaced by the

Laplace-Beltrami operator in the new curvilinear coordinates $q = q(\bar{q})$, so the solution is not changed, but written in the new coordinates. This is similar to the covariance of the ordinary Schrödinger equation and the corresponding *Lagrangian* path integral relative to general coordinate transformations on the configuration space: The kinetic energy operator (the Laplace operator) in the ordinary Schrödinger equation provides a term *quadratic* in time derivatives in the path integral measure which is sufficient for the general coordinate covariance. We remark that the regularization procedure based on the modified Schrödinger equation (A.50) applies to far more general Hamiltonians than those quadratic in canonical momenta and leading to the conventional *Lagrangian* path integral.

A.3.3 Modern String Actions and Transition Amplitudes

For the sake of completeness, in this subsection we give a brief review of modern path integral methods is quantum fields and string theory (mainly following [DEF99]). Recall that the fundamental quantities in quantum field theory (QFT) are the transition amplitudes, $IN \Longrightarrow OUT$, for processes in which a number of IN incoming particles scatter to produce a number of OUT outgoing particles. The square modulus of the transition amplitude yields the *probability* for this process to take place.

The only way we have today to define string theory is by giving a *rule* for the evaluation of transition amplitudes, order by order in the loop expansion, i.e., genus by genus. The rule is to assign a relative weight to a given configuration and then to sum over all configurations. To make this more precise, we first describe the system's configuration manifold M.

We assume that Σ and M are smooth manifolds, of dimensions 2 and n respectively, and that x is a continuous map from Σ to M. If ξ^m, $m = 1, 2$, are local coordinates on Σ and x^μ, $\mu = 1, \ldots, n$, are local coordinates on M then the map x may be described by functions $x^\mu(\xi^m)$ which are continuous.

To each system configuration we can associate a weight $\mathrm{e}^{-S[x,\Sigma,M]}$, $(S \in \mathbb{C})$ and the *transition amplitude*, Amp, for specified external strings (incoming and outgoing) is obtained by summing over all surfaces Σ and all possible maps x.

$$Amp = \sum_{\text{surfaces } \Sigma} \sum_{x} \mathrm{e}^{-S[x,\Sigma,M]} .$$

We now need to specify each of these ingredients:

1) We assume M to be an nD Riemannian manifold, with metric g. A special case is flat Euclidean space–time $\mathbb{R}^n$. The space–time metric is assumed *fixed*.

$$ds^2 = (dx, dx)_g = g_{\mu\nu}(x) dx^\mu \otimes dx^\nu .$$

2) The metric g on M induces a metric on Σ: $\gamma = x^*(g)$,

$$\gamma = \gamma_{mn} d\xi^m \otimes d\xi^n , \qquad \gamma_{mn} = g_{\mu\nu} \frac{\partial x^\mu}{\partial \xi^m} \frac{\partial x^\nu}{\partial \xi^n} .$$

This metric is non–negative, but depends upon x. It is advantageous to introduce an intrinsic Riemannian metric g on Σ, independently of x; in local coordinates, we have

$$g = g_{mn}(\xi) d\xi^m \otimes d\xi^n.$$

A natural intrinsic candidate for S is the area of $x(\Sigma)$, which gives the *Nambu–Goto action*

$$\text{Area } (x(\Sigma)) = \int_\Sigma d\mu_\gamma = \int_\Sigma n^2 \xi \sqrt{\det \gamma_{mn}},$$

which depends only upon g and x, but not on g. However, the transition amplitudes derived from the Nambu–Goto action are *not well–defined* quantum–mechanically.

Otherwise, we can take as starting point the *Polyakov action*

$$S[x, g] = \kappa \int_\Sigma (dx, *dx)_g = \kappa \int_\Sigma d\mu_g g^{mn} \partial_m x^\mu \partial_n x^\nu g_{\mu\nu}(x),$$

where κ is the *string tension* (a positive constant with dimension of inverse length square). The stationary points of S with respect to g are at $g^0 = \mathrm{e}^\phi \gamma$ for some function ϕ on Σ, and thus $S[x, g^0] \sim \text{Area } (x(\Sigma))$.

The Polyakov action leads to *well–defined* transition amplitudes, obtained by integration over the space $\text{Met}(\Sigma)$ of all positive metrics on Σ for a given topology, as well as over the space of all maps $\text{Map}(\Sigma, M)$. We can define the path integral

$$Amp = \sum_{\substack{\text{topologies} \\ \Sigma}} \int_{\text{Met}(\Sigma)} \frac{1}{N(g)} \int_{\text{Map}(\Sigma, M)} \mathcal{D}[x] \, \mathrm{e}^{-S[x,g,g]},$$

where N is a normalization factor, while the measures $\mathcal{D}[g]$ and $\mathcal{D}[x]$ are constructed from $\text{Diff}^+(\Sigma)$ and $\text{Diff}(M)$ invariant L^2 norms on Σ and M. For fixed metric g, the action S is well–known: its stationary points are the harmonic maps $x\colon\ \Sigma \to M$. Here, however, g varies and in fact is to be integrated over. For a general metric g, the action S defines a *nonlinear sigma model*, which is renormalizable because the dimension of Σ is 2. It would not in general be renormalizable in dimension higher than 2, which is usually regarded as an argument against the existence of fundamental membrane theories (see [DEF99]).

Transition Amplitude for a Single Point Particle

The transition amplitude for a single point particle could in fact be obtained in a way analogous to how we prescribed string amplitudes. Let space–time be again a Riemannian manifold M, with metric g. The prescription for the transition amplitude of a particle travelling from a point $y \in M$ to a point y'

to M is expressible in terms of a sum over all (continuous) paths connecting y to y':

$$Amp(y,y') = \sum_{\substack{\text{paths} \\ \text{joining } y \text{ and } y'}} \mathrm{e}^{-S[\text{path}]}.$$

Paths may be parametrized by maps from $C = [0,1]$ into M with $x(0) = y$, $x(1) = y'$. A simple worldline action for a massless particle is obtained by introducing a metric g on $[0,1]$

$$S[x,g] = \frac{1}{2}\int_C d\tau\, g(\tau)^{-1}\dot{x}^\mu \dot{x}^\nu g_{\mu\nu}(x),$$

which is invariant under $\mathrm{Diff}^+(C)$ and $\mathrm{Diff}(M)$.

Recall that the analogous prescription for the point particle transition amplitude is the path integral

$$Amp(y,y') = \int\limits_{\mathrm{Met}(C)} \mathcal{D}[g]\frac{1}{N}\int\limits_{\mathrm{Map}(C,M)} \mathcal{D}[x]\,\mathrm{e}^{-S[x,g]}.$$

Note that for string theory, we had a prescription for transition amplitudes valid for all topologies of the worldsheet. For point particles, there is only the topology of the interval C, and we can only describe a single point particle, but not interactions with other point particles. To put those in, we would have to supply additional information.

Finally, it is very instructive to work out the amplitude Amp by carrying out the integrations. The only $\mathrm{Diff}^+(C)$ invariant of g is the length $L = \int_0^1 d\tau\, g(\tau)$; all else is generated by $\mathrm{Diff}^+(C)$. Defining the normalization factor to be the volume of $\mathrm{Diff}(C)$: $N = \mathrm{Vol}(\mathrm{Diff}(C))$ we have $\mathcal{D}[g] = \mathcal{D}[v]\,dL$ and the transition amplitude becomes

$$Amp(y,y') = \int_0^\infty dL \int \mathcal{D}[x]\,\mathrm{e}^{-\frac{1}{2L}\int_0^1 d\tau(\dot{x},\dot{x})_g} = \int_0^\infty dL\,\langle y'|\mathrm{e}^{-L\Delta}|y\rangle = \left\langle y'\right|\frac{1}{\Delta}\left|y\right\rangle.$$

Thus, the amplitude is just the Green function at (y,y') for the Laplacian Δ and corresponds to the propagation of a massless particle (see [DEF99]).