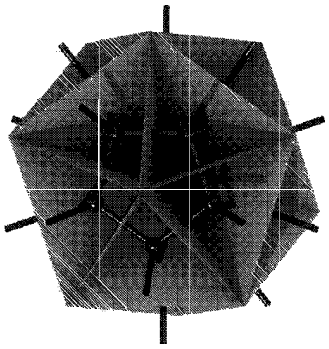
Quantum Gravity

Carlo Rovelli



DRAFT

Version: December 30, 2003

Contents

Preface

A dream I have long held was to write a "treatise" on quantum gravity once the theory had been finally found and experimentally confirmed. We are not yet there. There isn't either experimental support nor large theoretical consensus. Still, a large amount of work has been developed over the last twenty years towards a quantum theory of spacetime. Many issues have been clarified, and a definite approach has crystallized. The approach, variously denoted[1], is mostly known as "loop quantum gravity".

The problem of quantum gravity has many aspects. Ideas and results are scattered in the literature. In this book I have attempted to collect the main results and to present an overall perspective on quantum gravity, as developed during this period. The point of view is personal and the choice of subjects is determined by my own interests. I apologize with friends and colleagues for what is missing: the reason so much is missing is in my own limits, for which I am the first to be sorry.

It is difficult to underestimate the vastitude of the problem of quantum gravity. The physics of the early XXth century has modified our understanding of the physical world in depth, changing the meaning of the basic concepts we use to grasp it: matter, causality, space and time. We haven't found a consistent picture of the world in which these modifications make sense together, yet. The problem of quantum gravity is nothing less than the problem of finding the novel consistent picture, finally bringing the XXth century scientific revolution to an end.

Solving a problem of this sort is not just a matter of mathematical skill. Like for the birth of quantum mechanics, relativity, electromagnetism, and newtonian mechanics, there are conceptual and foundational problems to be addressed. We have to understand which (possibly new) notions make sense and which old notions must be discarded, in order to describe spacetime in the quantum relativistic regime. What we need is not just a technique for computing, say, graviton-graviton scattering amplitudes (although we certainly want to be able to do so, eventually). What we need is to understand how to think the world at the light of what we have learned about it with quantum theory and general relativity.

General relativity, in particular, has modified our understanding of the spatio-temporal structure of reality in a way whose consequences have not been fully explored yet. A consistent part of the research in quantum gravity explores foundational issues, and Part I of this book ("Relativistic foundations") is devoted to basic issues. It is an exploration on how to rethink basic physics from scratch, after the general-relativistic conceptual revolution. Without this, we risk to ask any tentative quantum theory of gravity the wrong kind of questions.

Part II of the book ("Loop quantum gravity") focuses on the loop approach. The loop theory, described in Part II, can be studied by itself, but its reason and interpretation are only clear in the light of the general framework studied in Part I. Although several aspects of this theory are still incomplete, the subject is mature for a book. A theory begins to be credible only when its

[1] See the notation section.

original predictions are reasonably unique and are confirmed by new experiments. Loop quantum gravity is not yet credible in this sense. Nor is any other current tentative theory of quantum gravity. The interest of the loop theory, in my opinion, is that at present it is the only approach to quantum gravity leading to well defined physical predictions (falsifiable, at least in principle) and, more importantly, it is the most determined effort for a genuine merge of quantum field theory with the world view that we have discovered with general relativity. The future will tell us more.

There are several other introductions to loop quantum gravity. Classic reports on the subject, illustrating various stages of the development of the theory are, in chronological order, [1, 2, 3, 4, 5, 6, 7, 8]. For a rapid orientation, and to appreciate different points of view, see the review papers [9, 10, 11, 12, 13]. Much useful material can be found in [14]. Good introductions to spin foam theory are in [9, 15, 16, 17]. This book is self contained, but I have tried to avoid excessive duplications, referring to other books and review papers for non-essential topics well developed elsewhere. This book focuses on physical and conceptual aspects of loop quantum gravity. Thomas Thiemann's book [18], which is going to be completed soon, focuses on the mathematical foundation of the same theory. The two books are complementary and can almost be read as Volume 1 ("Introduction and conceptual framework") and Volume 2 ("Complete mathematical framework") of a general presentation of loop quantum gravity.

The book assumes that the reader has a basic knowledge of general relativity, quantum mechanics and quantum field theory. In particular, the aim of the chapters on general relativity (chapter 2), classical mechanics (ch 3), hamiltonian general relativity (ch 4), and quantum theory (ch 5) is to offer the fresh perspective on these topics which is needed for quantum gravity, to a reader that already knows the conventional formulation of these theories.

Sections with comments and examples are printed in smaller fonts. Sections that contain side or more complex topics and that can be skipped in a first reading without compromising the understanding of what follows are marked with a star (*) in the title. References in the text appear only when strictly needed for comprehension. Each chapter ends with a short bibliographical section, pointing out essential references for the reader who wants to go more in detail or to trace original works on a subject. I have given up the immense task of collecting a full bibliography on loop quantum gravity. On many topics I refer to specific review articles where ample bibliographic information can be found. An extensive bibliography on loop quantum gravity is in [18].

I have written this book thinking of a researcher interested in working in quantum gravity, but also of a good PhD student or an open minded scholar, curious about this extraordinary open problem. I have found the journey towards general relativistic quantum physics, towards quantum spacetime, a fascinating adventure. I hope the reader will see the beauty I see, and that he or she will be capable to complete the journey. The landscape is magic, the trip is far from being over.

Marseille, Toronto, Rome, 2002-2003

Acknowledgements

I am indebted with the many people that have sent suggestions and corrections to the draft of this book posted online. Special thanks in particular to Justin Malecki, Simone Speziale, Luisa Doplicher and Leonard Cottrell.

My great gratitude to the friends with whom I have had the privilege of sharing this adventure:

To Lee Smolin, companion of adventures and friend. His unique creativity and intelligence, intellectual freedom and total honesty, are among the very best things I have found in life.

To Abhay Ashtekar whose tireless analytical rigor, synthesis capacity and leadership have been a most precious guide. Abhay has solidified our ideas and transformed our intuitions in theorems. This book is a result of Lee's and Abhay's ideas and work as much as my own.

To Laura Scodellari and Chris Isham, my first masters who have guided me into mathematics and quantum gravity.

To Ted Newman, who, with Sally, has parented the little boy just arrived from the Empire's far provinces. I have shared with Ted a decade of intellectual joy. His humanity, generosity, honesty, passion and love for thinking, are the example against which I judge myself.

To Laura Doplicher Simone Speziale, Florian Conrady, Daniele Colosi, Etera Livine, Daniele Oriti, Florian Girelli, Roberto DePietri, Robert Oeckl, Merced Montesinos, Kirill Krasnov, Carlos Kozameh, Michael Reisenberger, Don Marolf, Berndt Brügmann, Hugo Morales-Tecotl, Laurent Freidel, Renate Loll, Alejandro Perez, Giorgio Immirzi, Philippe Roche, Federico Laudisa, Jorge Pullin, Thomas Thiemann, Louis Crane, Jerzy Lewandowski, John Baez, Ted Jacobson, Marco Toller, Jeremy Butterfield, John Norton, John Barrett, Jonathan Halliwell, Massimo Testa, David Finkelstein, Gary Horowitz, John Earman, Julian Barbour, John Stachel, Massimo Pauri, Jim Hartle, Roger Penrose, John Wheeler, Alain Connes, and all the many other friends with whom the ideas and results described in this book have been developed.

With all these friends I have had the joy of talking about a physics which is far from problem-solving, from outsmarting each other, or from making weapons to make us stronger than them. I think that physics is about escaping the prison of the received thoughts and searching for novel ways of thinking the world, about trying to clear a bit the misty lake of our insubstantial dreams, which reflect reality like the lake reflects the mountains.

Foremost, thanks to Bonnie, she knows why.

Terminology and notation

In the book, "relativistic" means "*general* relativistic", unless otherwise specified. When referring to *special* relativity, I say so explicitly. Similarly, "nonrelativistic" and "prerelativistic" mean "non *general* relativistic" and "pre-*general*-relativistic". The choice is a bit unusual (special relativity, in this language, is "nonrelativistic"). One reason for it is simply to make language smoother: the book is about *general* relativistic physics, and repeating "*general*" every other line sounds too much like a Frenchman talking about de Gaulle. But there is a more substantial reason: the complete revolution in spacetime physics, which truly deserves the name of relativity is the one of general relativity, not the one of special relativity. This opinion is not always shared today, but it was Einstein's opinion. Einstein has been criticized on this; but in my opinion the criticisms miss the full reach of Einstein's discovery about spacetime. One of the aims of this book is to defend in modern language Einstein's intuition that his gravitational theory is the full implementation of relativity in physics. This point is discussed at length in chapter 2.

I often indulge to the bad physicists' habit of mixing up function (f) and function values ($f(x)$). Care is used when relevant. Similarly, I follow standard physicists abuse of language in denoting a field such as the Maxwell potential as $A_\mu(x)$, $A(x)$, or A, where the three notations are treated as equivalent manners of denoting the field. Again, care is used where relevant.

All fields are assumed to be smooth, unless otherwise specified. All statements about manifolds and functions are local unless otherwise specified; that is, they hold within a single coordinate patch. In general I do not specify the domain of definition of functions: clearly equations hold where functions are defined.

Index notation follows the most common choice in the field: Greek indices from the middle of the alphabet $\mu, \nu, \ldots = 0, 1, 2, 3$ are 4d spacetime tangent indices. Capital Latin indices from the middle of the alphabet $I, J, \ldots = 0, 1, 2, 3$ are 4d Lorentz tangent indices. (In the special relativistic context the two are used without distinction.) Lowercase Latin indices from the beginning of the alphabet $a, b, \ldots = 1, 2, 3$ are 3d tangent indices. Lowercase Latin indices from the middle of the alphabet $i, j, \ldots = 0, 1, 2, 3$ are 3d indices in R^3. Coordinates of a 4d manifold are usually indicated as $x, y \ldots$, while 3d manifold coordinates are usually indicated as $\vec{x}, \vec{y}$ (also as $\vec{r}$). Thus the components of a spacetime coordinate x are

$$x^\mu = (t, \vec{x}) = (x^0, x^a); \tag{1}$$

while the components of a Lorentz vector e are

$$e^I = (e^0, e^i). \tag{2}$$

η_{IJ} is the Minkowski metric, with signature $[-, +, +, +]$. The indices $I, J \ldots$ are raised and lowered with η_{IJ}. δ_{ij} is the Kronecker delta, or the R^3 metric. The indices $i, j \ldots$ are raised and lowered with δ_{ij}.

For reasons explained at the beginning of chapter 2, I call "gravitational field" the tetrad field $e^I_\mu(x)$, instead of the metric tensor $g_{\mu\nu}(x) = \eta_{IJ}\, e^I_\mu(x) e^J_\nu(x)$.

ϵ_{IJKL}, or $\epsilon_{\mu\nu\rho\sigma}$, is the completely antisymmetric object with $\epsilon_{0123} = 1$. Same for ϵ_{abc}, or ϵ_{ijk}, in 3d. The Hodge star is defined by

$$F^*_{IJ} = \epsilon_{IJKL} F^{KL} \tag{3}$$

in flat space, and by the same equation, where $F_{IJ} e^I_\mu e^J_\nu = F_{\mu\nu}$ and $F^*_{IJ} e^I_\mu e^J_\nu = F^*_{\mu\nu}$ in the presence of gravity. Equivalently,

$$F^*_{\mu\nu} = \sqrt{-\det g}\ \epsilon_{\mu\nu\rho\sigma} F^{\rho\sigma} = |\det e|\ \epsilon_{\mu\nu\rho\sigma} F^{\rho\sigma}. \tag{4}$$

Symmetrization and antisymmetrization of indices is defined with a half: $A_{(ab)} = \frac{1}{2}(A_{ab} + A_{ba})$ and $A_{[ab]} = \frac{1}{2}(A_{ab} - A_{ba})$.

I call "curve" on a manifold M, a map

$$\begin{aligned} \gamma : I &\to M \\ s &\mapsto \gamma^a(s), \end{aligned} \tag{5}$$

where I is an interval of the real line R (possibly the entire R.) I call "path" an oriented unparametrized curve, namely an equivalence class of curves under change of parametrization $\gamma^a(s) \mapsto \gamma'^a(s) = \gamma^a(s'(s))$, with $ds'/ds > 0$.

An orthonormal basis in the Lie algebras $su(2)$ and $so(3)$ is chosen once and for all and these algebras are identified with R^3. For $so(3)$, the basis vectors $(v_i)^j{}_k$ can be taken proportional to $\epsilon_i{}^j{}_k$; for $su(2)$, the basis vectors $(v_i)^A{}_B$ can be taken proportional to the Pauli matrices, recalled in the appendix. Thus, an algebra element ω in $su(2) \sim so(3)$ has components ω^i.

For any antisymmetric quantity v^{ij} with two 3d indices i, j, I use also the one-index notation

$$v^i = \tfrac{1}{2}\, \epsilon^i{}_{jk}\ v^{jk}, \qquad v^{ij} = \epsilon^{ij}{}_k\ v^k; \tag{6}$$

the one-index and the two-indices notation are considered defining the same object. For instance the $SO(3)$ connections ω^{ij} and A^{ij}, are equivalently denoted ω^i and A^i.

Symbols. Here is a list of symbols, with their name and the equation or the section where they are introduced or defined.

$\mathbf{A}$	area	Sec 2.1.4
A	Yang-Mills connection	Eq 2.27
$A, A^i_\mu(x)$	selfdual 4d gravitational connection	Eq 2.16
$A, A^i_a(\vec{x})$	selfdual or real 3d gravitat. conn.	Sec 4.1.1, 4.2
$\mathcal{C}$	relativistic configuration space	Sec 3.2.1
D_μ	covariant derivative	Eq 2.28
$Diff^*$	extended diffeomorphism group	Sec 6.2.2
$e^I_\mu(x)$	gravitational field	Eq 2.1
e	determinant of e^I_μ	
e	edge (of spinfoam)	Sec 9.1
$E, E^a_i(\vec{x})$	gravitational electric field	Sec 4.1.1

f	face (of spinfoam)	Sec 9.1
F	curvature two-form	Sec 2.1.1
g or U	group element	
G	Newton constant	
$\mathcal{G}$	space of boundary data	Sec 3.2.5-3.3.3
h_γ	$U(A,\gamma)$	Sec 7.1
H	relativistic hamiltonian	Sec 3.2
H_0	nonrelativistic (conventional) hamilt	Sec 3.2
$\mathcal{H}$	quantum state space	Ch 5
$\mathcal{H}_0$	nonrelativistic quantum state space	Ch 5
i_n	intertwiner on spinnetwork node n	Sec 6.3
i_e	intertwiner on spinfoam edge e	Ch 9
j	irreducible rep (for $SU(2)$: spin)	
j_l	spin associated to spinnetwork link l	Sec 6.2.1
j_f	rep associated to spinfoam face f	Ch 9
$\mathcal{K}$	kinematical quantum state space	Sec 5.2
$\mathcal{K}_0$	$SU(2)$ invariant quantum state space	Sec 6.2.3
$\mathcal{K}_{\text{Diff}}$	diff invariant quantum state space	Sec 6.2.3
K	boundary quantum space	Sec 5.1.4 5.3.5
l	link (of spin network)	Sec 9.1
$\mathbf{L}$	length	Sec 2.1.4
M	spacetime manifold	
n	node (of spin network)	Sec 9.1
p_a	relativistic momenta (including p_t)	Sec 3.2
p_t	momentum conjugate to t	Sec 3.2
P	the "projector" operator	Sec 5.2
$P_{_G}$	group G projector	Eq 9.117
$P_{_H}$	subgroup H projector	Eq 9.119
$\mathcal{P}$	transition probability	Ch 5
q^a	partial observables	Sec 3.2
$R^I{}_{J\mu\nu}(x)$	curvature	Eq 2.8
$R^{(j)\alpha}{}_\beta(g)$	matrix of group element g in repr j	
$\mathcal{R}$	3d region	Sec 2.1.4
s	s-knot : abstract spin network	Eq 6.4.1
$\|s\rangle$	s-knot state	Eq 6.4.1
S_{BH}	black hole entropy	Sec 8.2
S	imbededd spin network	Sec 6.3
$\|\mathrm{S}\rangle$	spin network state	Sec 6.3.1
$\mathcal{S}$	2d surface	Sec 2.1.4
$\mathcal{S}$	space of fast decrease functions	Ch 5
$\mathcal{S}_0$	space of tempered distributions	Ch 5
$S[\tilde{\gamma}]$	action functional	Sec 3.2
$S(q^a)$	Hamilton-Jacobi function	Sec 3.2.2
$S(q^a, q_0^a)$	Hamilton function	Sec 3.2.5
t_ρ	thermal time	Sec 3.4 and 5.5.1
T	target space of a field theory	Sec 3.3.1
U or g	group element	
$U(A,\gamma)$	holonomy	Sec 2.1.5
v	vertex (of spinfoam)	Sec 9.1
$\mathbf{V}$	volume	Sec 2.1.4

$W(q^a, q'^a)$	propagator	Ch 5
W	transition amplitudes, propagator	Sec 5.2
x	4d spacetime coordinates	
$\vec{x}$	3d coordinates	
Z	partition function	Ch 9
α	loop, closed path	
β	inverse temperature	Sec 3.4
γ	path	
γ	motion (in $\mathcal{C}$)	Sec 3.2.1
γ	Immirzi parameter	Sec 4.2.3
$\tilde{\gamma}$	motion in Ω	Sec 3.2
Γ	relativistic phase space	Sec 3.2.1
Γ	graph	Ch 6.2
Γ	two-complex	Ch 9
θ	Poincaré-Cartan form on Σ	Sec 3.2.2
$\tilde{\theta}$	Poincaré form on Ω	Eq 3.9
$\eta_{\mu\nu}$	Minkowski metric = diag[-1,1,1,1]	
λ	cosmological constant	Eq 2.9
λ	gauge parameter	Sec 2.1.3
ρ	statistical state	Sec 3.4 and 5.5.1
Σ	constraint surface $H = 0$	Sec 3.2.2
Σ	3d boundary surface	Ch 4
σ	spinfoam	Ch 9
$\phi(x)$	scalar field	Eq 2.29
$\psi(x)$	fermion field	Eq 2.32
ω	presymplectic form on Σ	Sec 3.2.2
$\omega^I_{\mu J}(x)$	spin connection	Eq 2.2
$\tilde{\omega}$	symplectic form on Ω	Sec 3.2.2
Ω	space of observables and momenta	Sec 3.2-3.3.2
$\{6j\}$	Wigner 6j symbol	Eq 9.33
$\{10j\}$	Wigner 10j symbol	Eq 9.103
$\{15j\}$	Wigner 15j symbol	Eq 9.56
$\|0\rangle$	covariant vacuum in K	Sec 5.1.4, 5.3.5
$\|0_t\rangle$	dynamical vacuum in $\mathcal{K}_t$	Sec 5.1.4, 5.3.2
$\|0_M\rangle$	Minkowski vacuum in $\mathcal{H}$	Sec 5.1.4, 5.3.1

The name of the theory. Finally, a word about the name of the quantum theory of gravity described in this book. The theory is known as *"loop quantum gravity"* (LQG), or sometimes *"loop gravity"* for short. However, the theory is also designated in the literature using a variety of other names. I list here these other names, and the variations of their use, for the benefit of the disoriented reader.

- *"Quantum Spin Dynamics"* (QSD) is used as a synonymous of LQG. Within LQG, it is sometimes used to designate in particular the dynamical aspects of the hamiltonian theory.

- *"Quantum geometry"* is sometimes used as a synonymous of LQG. Within the theory, it is used to designate in particular the kinematical aspects of the theory. The expression *"quantum geometry"* is generic: it is also widely used in other approaches to quantum spacetime, in particular, dynamical

triangulations [19] and noncommutative geometry. Therefore it is not a good designation for a specific theory of quantum gravity.

- *"Nonperturbative quantum gravity"*, *"canonical quantum gravity"* and *"quantum general relativity"* (QGR) are often used to designate LQG, although their proper meaning is wider.

- The expression *"Ashtekar approach"* was used in the past to designate LQG: it comes from the fact that a key ingredient of LQG is the reformulation of classical GR as a theory of connections, developed in particular by Abhay Ashtekar.

- In the past, LQG was also called *"the loop representation of quantum general relativity"*. Today, *"loop representation"* and *"connection representation"* are used within LQG to designate representation of the states of LQG as functionals of loops (or spin networks) and, respectively, functionals of the connection. The two are related in the same manner as the energy ($\psi_n = \langle n|\psi\rangle$) and position ($\psi(x) = \langle x|\psi\rangle$) representations of the harmonic oscillator states.

Part I

Relativistic foundations

I know that I am mortal, and the creature of a day... but when I search out the massed wheeling circles of the stars, my feet no longer touch the earth: side by side with Zeus himself, I drink my fill of ambrosia, food of the gods...

Claudius Ptolemy, "Mathematical Syntaxis"

Chapter 1

General ideas and heuristic picture

The aim of this chapter is to introduce the general ideas on which this book is based and to present the picture of quantum spacetime that emerges from loop quantum gravity, in a heuristic and intuitive manner. The style of the chapter is therefore conversational, with little regard for precision and completeness. In the course of the book the ideas and notions introduced here will be made precise, and the claims will be justified and formally derived.

1.1 The problem of quantum gravity

1.1.1 Unfinished revolution

Quantum mechanics (QM) and general relativity (GR) have extended our understanding of the physical world widely. A large part of the physics of the last century has been a triumphant march of exploration of new worlds opened by these two theories. QM led to atomic physics, nuclear physics, particle physics, condensed matter physics, semiconductors, lasers, computers, quantum optics ... GR led to relativistic astrophysics, cosmology, GPS technology ... and is today leading us, hopefully, towards gravitational wave astronomy.

But QM and GR have destroyed the coherent picture of the world provided by prerelativistic classical physics: each was formulated in terms of assumptions contradicted by the other theory. QM was formulated using an external time variable (the t of the Schrödinger equation) or a fixed, nondynamical background spacetime (the spacetime on which quantum field theory is defined). But this external time variable and this fixed background spacetime are incompatible with GR. In turn, GR was formulated in terms of Riemannian geometry, assuming that the metric is a smooth and deterministic dynamical field. But QM requires that any dynamical field is quantized: at small scales it manifests itself in discrete quanta and is governed by probabilistic laws.

We have learned from GR that spacetime is dynamical and we have learned from QM that any dynamical entity is made by quanta and can be in probabilistic superposition states. Therefore at small scales there should be quanta of space and quanta of time, and quantum superposition of spaces. But what does this mean? We live in a spacetime with quantum properties: a *quantum spacetime*. What is quantum spacetime? How can we describe it?

Classical prerelativistic physics provided a coherent picture of the physical world. This was based on clear notions such as *time, space, matter, particles, waves, forces, measurements, deterministic laws*... This picture has partially evolved (in particular with the advent of field theory and special relativity) but it has remained consistent and quite stable for three centuries. GR and QM have then modified the basic notions in depth. GR has modified the notions of space and time; QM the notions of causality, matter, and measurements. The novel, modified notions do not fit together easily. The new coherent picture is not yet available. With all their immense empirical success,

GR and QM have left us with an understanding of the physical world which is unclear and badly fragmented. At the foundations of physics there is today confusion and incoherence.

We want to combine what we have learnt about our world from the two theories and to find a new synthesis. This is a major challenge –perhaps the major challenge– in today's fundamental physics. GR and QM have opened a revolution. The revolution is not yet complete.

With notable exceptions (Dirac, Feynman, Weinberg, DeWitt, Wheeler, Penrose, Hawking, t'Hooft, among others) most of the physicists of the second half of last century have ignored this challenge. The urgency was to apply the two theories to larger and larger domains. The developments were momentous and the dominant attitude was pragmatic. Applying the new theories was more important than understanding them. But an overly pragmatic attitude is not productive in the long run. Towards the end of the XXth century, the attention of theoretical physics has been increasingly focusing on the problem of merging the conceptual novelties of QM and GR.

This book is the account of an efforts to do so.

1.1.2 How to search for quantum gravity?

How to search for this new synthesis? Conventional field quantization methods are based on the weak field perturbation expansion. Their application to GR fails because it yields a nonrenormalizable theory. Perhaps this is not surprising: GR has changed the notions of space and time too radically, to docilely agree with flat space quantum field theory. Something else is needed.

In science there are no secure recipes for discovery and it is important to explore different directions at the same time. Currently, a quantum theory of gravity is sought along various paths. The two most developed are loop quantum gravity, described in this book, and string theory. Other research directions include dynamical triangulations, noncommutative geometry, Hartle's quantum mechanics of spacetime (this is not really a specific quantum theory of gravity, but rather a general theoretical framework for general relativistic quantum theory), Hawking's euclidean sum over geometries, quantum Regge calculus, Penrose's twistor theory, Sorking's causal sets, t'Hooft deterministic approach and Finkelstein's theory. The reader can find ample references in the general introductions to quantum gravity mentioned in the note at the end of this chapter. Here, I sketch only the general ideas that motivate the approach described in this book, plus a brief comment on string theory, which is the most popular alternative to loop gravity.

Our present knowledge of the basic structure of the physical universe is summarized by GR, quantum theory and quantum field theory (QFT), and the particle physics standard model. This set of fundamental theories is inconsistent. But it is characterized by an extraordinary empirical success, nearly unique in the history of science. Indeed, currently there is no evidence of any observed phenomenon that clearly escapes, questions or contradicts this set of theories (or a minor modification of the same, to account, say, for a neutrino mass or a cosmological constant). This set of theories becomes meaningless in certain physical regimes. In these regimes, we expect the predictions of quantum gravity to become relevant and to differ from the predictions of GR and the standard model. But these regimes are outside our experimental or observational reach, at least so far. Therefore, we have no direct empirical guidance for searching for quantum gravity – as, say, atomic spectra guided the discovery of quantum theory.

Since quantum gravity is a theory expected to describe regimes that are, so far, inaccessible, one might worry that anything could happen in these regimes, at scales far removed from our experience. Maybe the search is impossible because the space of the possible theories is too large. This worry is unjustified. If this was the problem, we would have plenty of complete, predictive and coherent theories of quantum gravity. Instead, the situation is precisely the opposite: we haven't any. The fact is that we do have plenty of information about quantum gravity, because we have QM and we have GR. Consistency with QM and GR is an extremely strict constraint.

A view is sometime expressed that some totally new, radical and wild hypothesis is needed for quantum gravity. I do not think that this is the case. Wild ideas pulled out of the blue sky have never made science advance. The radical hypotheses that physics has successfully adopted have always been reluctantly adopted because they were forced by new empirical data –Kepler's ellipses, Bohr's quantization ... – or by stringent theoretical deductions –Maxwell inductive current, Einstein's relativity... (See appendix C). Generally, arbitrary novel hypotheses lead nowhere.

In fact, today we are precisely in one of the typical situations in which theoretical physics has worked at its best in the past. Many of the most striking advances in theoretical physics have derived from the effort of finding a common theoretical framework for two basic and apparently conflicting discoveries. For instance, the aim of combining special relativity and non relativistic quantum theory led to the theoretical discovery of antiparticles; combining special relativity with Newtonian gravity led to general relativity; combining the Keplerian orbits with Galilean physics led to Newton's mechanics; combining Maxwell theory with Galilean relativity led to special relativity, and so on. In all these cases, major advances have been obtained by "taking seriously"[1] apparently conflicting theories, and exploring the implications of holding the key tenets of both theories for true. Today we are precisely in one of these characteristic situations. We have learned two new very general "facts" about nature, expressed by QM and GR: we have "just" to figure out what they imply, taken together. Therefore, the question we have to ask is: what have we really learned about the world from QM and from GR? Can we combine these insights into a coherent picture? What we need is a conceptual scheme in which the insights obtained with GR and QM fit together.

This view is *not* the majority view in theoretical physics, at present. There is consensus that QM has been a conceptual revolution, but many do not view GR in the same way. According to many, the discovery of GR has been just the writing of one more field theory. This field theory is, furthermore, likely to be only an approximation to a theory we do not yet know. According to this opinion, GR should not be taken too seriously as a guidance for theoretical developments.

I think that this opinion derives from a confusion: the confusion between the specific form of the Einstein-Hilbert action and the modification of the notions of space and time engendered by GR. The Einstein-Hilbert action might very well be a low energy approximation of something else. But the modification of the notions of space and time has to do with the diffeomorphism invariance and the background independence of the action, not with its specific form. If we make this confusion, we underestimate the novelty of the physical content of GR. The challenge of quantum gravity is precisely to fully incorporate this radical novelty into QFT. In other words, the task is to understand what is a general relativistic QFT, or a background independent QFT.

Today many physicists prefer disregarding or postponing these foundational issues and, instead, develop and adjust current theories. The most popular strategy towards quantum gravity, in particular, is to pursue the line of research grown in the wake of the success of the standard model of particle physics. The failure of perturbative quantum GR is interpreted as a replay of the failure of Fermi theory.[2] Namely as an indication that we must modify GR at high energies. With the input of the grand-unified-theories, supersymmetry, and Kaluza-Klein theory, the search for a high energy correction of GR free from bad ultraviolet divergences has led to higher derivative theories, supergravity, and finally to string theory.

Sometimes the claim is made that the quantum theory of gravity has already been found and it is string theory. Since this is a book about quantum gravity without strings, I should say a few words about this claim. String theory is based on a physical hypothesis: elementary objects are extended,

---

[1] In [20], Gell-Mann says that the main lesson to be learnt by from Einstein is "to 'take very seriously' ideas that work, and see if they can be usefully carried much further than the original proponent suggested".

[2] Fermi theory was an empirically successful but nonrenormalizable theory of the weak interactions, like GR is an empirically successful but nonrenormalizable theory of the gravitational interaction. The solution has been the Glashow-Weinberg-Salam electroweak theory, which corrects Fermi theory at high energy.

rather than particle-like. This hypothesis leads to a very rich unified theory, which contains much phenomenology, including (with suitable inputs) fermions, Yang-Mills fields and gravitons, and is believed by many to be free of ultraviolet divergences. The price to pay for these theoretical results is a gigantic baggage of additional physics: supersymmetry, extra dimensions, an infinite number of fields with arbitrary masses and spins, and so on.

So far, nothing of this new physics shows up in experiments. Supersymmetry, in particular, has been claimed to be on the verge of being discovered for years, but hasn't shown up. Unfortunately so far the theory can accommodate any disappointing experimental result because it is hard to derive precise new quantitative physical predictions, with which the theory could be falsified, from the monumental mathematical apparatus of the theory. Furthermore, even recovering the real world is not easy within the theory: the search for a compactification leading to the standard model, with its families and masses and no instabilities, has not yet succeeded, as far as I know. It is clear that string theory is a very interesting hypothesis, but certainly not an established theory. It is therefore important to pursue alternative directions as well.

String theory is a direct development of the standard model and is deeply rooted in the techniques and the conceptual framework of flat space QFT. As I shall discuss in detail all along this book, many of the tools used in this framework – energy, unitary time evolution, vacuum state, Poincaré invariance, S-matrix, objects moving in a spacetime, Fourier transform ... – do not make sense anymore in the quantum gravitational regime, in which the gravitational field cannot be approximated by a background spacetime – perhaps not even asymptotically.[3] Therefore string theory does not address directly the main challenge of quantum gravity: understanding what is background independent QFT. Facing this challenge directly, before worrying about unification, leads, instead, to the direction of research investigated by loop quantum gravity.[4]

The alternative to the line of research followed by string theory is given by the possibility that the failure of perturbative quantum GR is *not* a replay of Fermi theory. That is, it is not due to a flaw of the GR action, but, instead, it is due to the fact that the conventional weak field quantum perturbation expansion cannot be applied to the gravitational field.

This possibility is strongly supported a posteriori by the results of loop quantum gravity. As we shall see, loop quantum gravity leads to a picture of the short scale structure of spacetime extremely different from that of a smooth background geometry. (There are hints in this direction from string theory calculations as well [23].) Spacetime turns out to have a nonperturbative, quantized, discrete structure at the Planck scale, which is explicitly described by the theory. The ultraviolet divergences may be cured by this structure. The ultraviolet divergences that appear in the perturbation expansion of conventional QFT may be a consequence of the fact that we erroneously replace this discrete Planck scale structure with a smooth background geometry.

If this is physically correct, ultraviolet divergences do not require the heavy machinery of string theory to be cured. On the other hand, the conventional weak field perturbative methods cannot be applied, because we cannot work with a fixed smooth background geometry. We must therefore adapt QFT to the full conceptual novelty of GR, and in particular to the change in the notion of space and time induced by GR. What are these changes? I sketch an answer below, leaving a

[3] To be sure, the development of string theory has incorporated many aspects of GR, such as curved spacetimes, horizons, black holes and relations between different backgrounds. But this is far from a background independent framework, such as the one realized by GR in the classical context. GR is not about physics on a curved spacetime, or about relations between different backgrounds: it is about the dynamics of spacetime. A background independent fundamental definition of string theory is being actively searched along several directions, but so far the definition of the theory and all calculations rely on background metric spaces.

[4] It has also been repeatedly suggested that loop gravity and string theory might merge, because loop gravity has developed precisely the background independent QFT methods that string theory needs [21]. Also, excitations over a weave (see section 6.7.1) have a natural string structure in loop gravity [22].

complete discussion to Chapter 2.

1.1.3 The physical meaning of general relativity

GR is the discovery that spacetime and the gravitational field are the same entity. What we call "spacetime" is itself a physical object, in many respects similar to the electromagnetic field. We can say that GR is the discovery that there is no spacetime at all. What Newton called "space", and Minkowski called "spacetime", is unmasked: it is nothing but a dynamical object –the gravitational field– in a regime in which we neglect its dynamics.

In Newtonian and special relativistic physics, if we take away the dynamical entities –particles and fields– what remains is space and time. In general relativistic physics, if we take away the dynamical entities, nothing remains. The space and time of Newton and Minkowski are reinterpreted as a configuration of one of the fields, the gravitational field. This implies that physical entities –particles and fields– are not all immersed in space, and moving in time. They do not live on spacetime. They live, so to say, on one another.

It is as if we had observed in the ocean many animals living on an island: animals on the island. Then we discover that the island itself is in fact a great whale. Not anymore animals on the island, just animals on animals. Similarly, the universe is not made by fields on spacetime; it is made by fields on fields. This book studies the far reaching effect that this conceptual shift has on QFT.

One consequence is that the quanta of the field cannot live in spacetime: they must build "spacetime" themselves. This is precisely what the quanta of space do in loop quantum gravity.

We may continue to use the expressions "space" and "time" to indicate aspects of the gravitational field, and I do so in the book. We are used to this in classical GR. But in the quantum theory, where the field has quantized "granular" properties and its dynamics is quantized and therefore only probabilistic, most of the "spatial" and "temporal" features of the gravitational field are lost.

Therefore for understanding the quantum gravitational field we must abandon some of the emphasis on geometry. Geometry represents well the classical gravitational field, not quantum spacetime. This is not a betrayal of Einstein's legacy: to the contrary, it is a step in the direction of "relativity" in the precise sense meant by Einstein. Alain Connes has beautifully described the existence of two points of view on space: the geometrical one, centered on the space points, and the algebraic, or "spectral" one, centered on the algebra of the dual spectral quantities. As emphasized by Alain, quantum theory forces us to a complete shift to this second point of view, because of noncommutativity. In the light of quantum theory, continuous spacetime cannot be anything else than an approximation in which we disregard quantum noncommutativity. In loop gravity, the physical features of space appear as spectral properties of quantum operators that describe our interactions with the gravitational field.

The key conceptual difficulty of quantum gravity is therefore to accept the idea that we can do physics in the absence of the familiar stage of space and time. We need to free ourselves from the prejudices associated with the habit of thinking of the world as "inhabiting space" and "evolving in time". Chapter 3 describes a general language for describing mechanical systems in this generalized conceptual framework.

1.1.4 Background independent quantum field theory

Is quantum mechanics[5] compatible with the general relativistic notions of space and time? It is, provided that we choose a sufficiently general formulation. For instance, the Schrödinger picture is

[5] I use the expression "quantum mechanics" to indicate the theory of all quantum systems, with a finite or infinite number of degrees of freedom. In this sense QFT is part of quantum mechanics.

only viable for theories where there is a a global observable time variable t; this conflicts with GR, where no such variable exists. Therefore the Schrödinger picture makes little sense in a background independent context. But there are formulations of quantum theory that are more general than the Schrödinger picture. In chapter 5, I describe a formulation of QM sufficiently general to deal with general relativistic systems. (For another relativistic formulation of QM, see [24].) Formulations of this kind are sometimes denoted "generalized quantum mechanics". I prefer denoting "quantum mechanics" any formulation of quantum theory, irrespectively of its generality, as "classical mechanics" is used to designate formalisms with different degrees of generality, such as Newton's, Lagrange's, Hamilton's or symplectic mechanics.

On the other hand, most of the conventional machinery of perturbative QFT is profoundly incompatible with the general relativistic framework. There are many reasons for this:

- The conventional formalism of QFT relies on Poincaré invariance. In particular, it relies on the notion of energy and on the existence of the nonvanishing hamiltonian operator that generates unitary time evolution. The vacuum, for instance, is the state that minimizes the energy. Generally, there is no global Poincaré invariance, no general notion of energy and no nonvanishing hamiltonian operator in a general relativistic theory.
- At the roots of conventional QFT is the physical notion of particle. The theoretical experience with QFT on curved spacetime [25] and on the relation between acceleration and temperature in QFT [26] indicates that in a generic gravitational situation the notion of particle can be quite delicate. (This point is discussed in Section 5.3.4.)
- Consider a conventional renormalized QFT. The physical content of the theory can be expressed in terms of its n-point functions $W(x_1, \ldots, x_n)$. The n-point functions reflect the invariances of the classical theory. In a general relativistic theory, invariance under a coordinate transformation $x \to x' = x'(x)$ implies immediately that the n-point functions must satisfy

$$W(x_1, \ldots, x_n) = W(x'(x_1), \ldots, x'(x_n)) \tag{1.1}$$

  and therefore (if the points in the argument are distinct) it must be a constant!

$$W(x_1, \ldots, x_n) = constant. \tag{1.2}$$

  Clearly we are immediately in a very different framework from conventional QFT.
- Similarly, the behavior for small $|x-y|$ of the two point function of a conventional QFT

$$W(x,y) = \frac{constant}{|x-y|^d}. \tag{1.3}$$

  expresses the short distance structure of the QFT. More generally, the short distance structure of the QFT is reflected in the operator product expansion

$$O(x)O'(y) = \sum_n \frac{O_n(x)}{|x-y|^n}. \tag{1.4}$$

  Here $|x-y|$ is the distance measured in the spacetime metric. On flat space for instance $|x-y|^2 = \eta_{\mu\nu}(x^\mu - y^\mu)(x^\nu - y^\nu)$. In a general relativistic context these expressions make no sense, since there is no background Minkowski (or other) metric $\eta_{\mu\nu}$. In its place, there is the gravitational field, namely the quantum field operator itself. But then, if standard operator product expansion becomes meaningless, the short distance structure of a quantum gravitational theory must be profoundly different from that of conventional QFT. As we shall see in Chapter 7 this is precisely the case.

There is a tentative escape strategy to circumvent these difficulties: write the gravitational field $e(x)$ as the sum of two terms

$$e(x) = e_{\text{background}}(x) + h(x); \tag{1.5}$$

where $e_{\text{background}}(x)$ is a background field configuration. This may be Minkowski, or any other. Assume that $e_{\text{background}}(x)$ defines spacetime, namely it defines location and causal relations. Then consider $h(x)$ as the gravitational field, governed by a QFT on the spacetime background defined by $e_{\text{background}}$. For instance the field operator $h(x)$ is assumed to commute at spacelike separations, where spacelike is defined in the geometry determined by $e_{\text{background}}(x)$. As a second step one may then consider conditions on $e_{\text{background}}(x)$ or relations between the formulations of the theory defined by different choices of $e_{\text{background}}(x)$. This escape strategy leads to three orders of difficulties: (i) Conventional perturbative QFT of GR based on (1.5) leads to a nonrenormalizable theory. To get rid of the uncontrollable ultraviolet divergences one has to get to the complications of string theory. (ii) As mentioned, loop quantum gravity shows that the structure of spacetime at the Planck scale is discrete. Therefore physical spacetime doesn't have a short distance structure at all. The unphysical assumption of a smooth background $e_{\text{background}}(x)$ implicit in (1.5) may be precisely the cause of the ultraviolet divergences. (iii) The separation of the gravitational field from spacetime is in strident contradiction with the very physical lesson of GR. If GR is of any guide in searching for a quantum theory of gravity, the relevant spacetime geometry is the one determined by the full gravitational field $e(x)$, and the separation (1.5) is misleading.

A formulation of quantum gravity that does not take the escape strategy (1.5) is a *background independent*, or general covariant QFT. The main aim of this book is develop the formalism for background independent QFT.

1.2 Loop quantum gravity

I sketch here the physical picture of quantum spacetime that emerges from loop quantum gravity (LQG). The basic ideas and assumptions on which LQG is based are the following.

(i) Quantum mechanics and general relativity. QM, suitably formulated to be compatible with general covariance, is assumed to be correct. The Einstein equations may be modified at high energy, but the general relativistic notions of space and time are assumed to be correct. The motivation for these two assumptions is the extraordinary empirical success they have had so far, and the absence of any contrary empirical evidence.

(ii) Background independence. LQG is based on the idea that the quantization strategy based on the separation (1.5) is *not* appropriate for describing the quantum properties of spacetime.

(iii) No unification. Nowadays, a fashionable idea is that the problem of quantizing gravity has to be solved together with the problem of finding a unified description of all interactions. LQG is a solution of the first problem, not the second.[6]

(iv) Four spacetime dimensions and no supersymmetry. LQG is compatible with these possibilities, but there is nothing in the theory that *requires* higher dimensions or supersymmetry. Higher spacetime dimensions and supersymmetry are interesting theoretical ideas, which, as many

[6]A motivation for the idea that these two issues are connected is the expectation that we are "near the end of physics". Unfortunately, the expectation of being "near the end of physics" has been present all along the three centuries of the history of modern physics. In the present situation of deep conceptual confusion on the fundamental aspects of the world, I see no sign indicating that we are close to the end of our discoveries about the physical world. When I was a student, it was fashionable to claim that the problem of finding a theory of the strong interactions had to be solved together with the problem of getting rid of renormalization theory. Nice idea. But wrong.

other interesting theoretical ideas, can be physically wrong. In spite of 15 years of search, numerous preliminary announcements of discovery then turned out out to be false, and repeated proclamations that supersymmetry was going to be discovered "next year", so far empirical evidence has been solidly and consistently *against* supersymmetry. This might change, but for the moment, as scientists, we must take the indications of the experiments seriously.

On the basis of these assumptions, LQG is a straightforward quantization of GR with its conventional matter couplings. The program of LQG is therefore conservative, and of small ambition. The physical inputs of the theory are just QM and GR, well tested physical theories. No major additional physical hypothesis or assumption is made (such as: elementary objects are strings, space is made by individual discrete points, quantum mechanics is wrong, GR is wrong, supersymmetry, extra dimensions...). No claim of being the final "Theory Of Everything" is made.

On the other hand, LQG has a radical and ambitious side: to merge the conceptual insight of GR into QM. In order to achieve this, we have to give up the familiar notions of space and time. The space continuum "on which" things are located and the time "along which" evolution happens are semiclassical approximate notions in the theory. In LQG, this radical step is assumed entirely.

The price of taking seriously the conceptual novelty of GR is that most of the traditional machinery of QFT becomes inadequate. This machinery is based on the existence of background spacetime, and GR is the discovery that there is no background spacetime. Therefore LQG does not make use of most of the familiar tools of conventional QFT; it only makes use of the general tools of quantum theory: a Hilbert space of states, operators related to the measurement of physical quantities, and transition amplitudes that determine the probability outcome of measurements of these quantities.

In LQG, Hilbert space of states and operators associated to physical observables are obtained from classical GR following a rather standard quantization strategy. A quantization strategy is a technique for searching a solution for a quite well posed inverse problem: finding a quantum theory with a given classical limit. The inverse problem could have many solutions. As noticed, presently the difficulty is not to discriminate among many complete and consistent quantum theories of gravity. We would be content with one.

1.2.1 Why loops?

Among the technical choices to make in order to implement a quantization procedure is the choice of the algebra of field functions to promote to quantum operators. In conventional QFT, this is generally the canonical algebra formed by the positive and negative components of the field modes. The quantization of this algebra leads to the creation and annihilation operators a and $a^\dagger$. The characterization of the positive and negative frequencies requires a background spacetime.

In contrast to this, what characterizes LQG is the choice of a different algebra of basic field functions: a non canonical algebra based on the holonomies of the gravitational connection. The holonomy (or "Wilson loop") is the parallel transport matrix along a closed curve.

The idea that holonomies are the natural variables in a gauge theory has a long history. In a sense, it can be traced back to the very origin of gauge theory, in the physical intuition of Faraday. Faraday understood electromagnetic phenomena in terms of "lines of force". Two key ideas underlie this intuition. First, that the relevant physical variables fill up space. This intuition by Faraday is the origin of field theory. Second, that the relevant variables do not refer to what happens at a point, but rather refer to the relation between different points connected by a line. The mathematical quantity that expresses this idea is the holonomy of the gauge potential along the line.

In LQG, the holonomy becomes a quantum operator that creates "loop states". In the loop representation formulation of Maxwell theory, for instance, a loop state $|\alpha\rangle$ is a state in which the electric field vanishes everywhere except along a single Faraday line α. More precisely, it is an

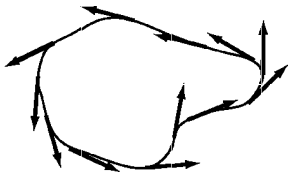


Figure 1.1: A loop α and the distributional electric field configuration $\vec{E}_\alpha$ (represented by the arrows).

eigenstate of the electric field with eigenvalue

$$\vec{E}_\alpha(\vec{x}) = \oint ds \; \frac{d\vec{\alpha}(s)}{ds} \; \delta^3(\vec{x}, \vec{\alpha}(s)), \tag{1.6}$$

where $s \mapsto \vec{\alpha}(s)$ is the Faraday line in space. This electric field vanishes everywhere except on the loop α itself, and at every point of α, it is tangent to the loop. See Figure (1.1). Notice that the vector distribution field $\vec{E}(\vec{x})$ defined in (1.6) is divergenceless, that is, it satisfies Coulomb law

$$\text{div } \vec{E}_\alpha(\vec{x}) = 0 \tag{1.7}$$

in the sense of distributions. In fact, for any smooth function f we have

$$\begin{aligned}
[\text{div } \vec{E}_\alpha](f) &= \int d^3x \; f(\vec{x}) \; \text{div } \vec{E}_\alpha(\vec{x}) \\
&= \int d^3x \; f(\vec{x}) \; \frac{\partial}{\partial x^a} \oint ds \; \frac{d\alpha^a(s)}{ds} \; \delta^3(\vec{x}, \vec{\alpha}(s)) \\
&= -\oint ds \; \frac{d\alpha^a(s)}{ds} \; \frac{\partial}{\partial \alpha^a} f(\alpha(s)) \\
&= -\oint_\alpha df \; = \; -\oint ds \; \frac{d}{ds} \; f(\alpha(s)) \; = \; 0.
\end{aligned} \tag{1.8}$$

Indeed, intuitively, Coulomb law requires precisely that an electric field at a point "continues" in the direction of the field itself, namely that it defines Faraday lines. The state $|\alpha\rangle$ is therefore a sort of minimal quantum excitation satisfying (1.7): it is an elementary quantum excitation of a single Faraday line.

The idea that a Yang-Mills theory is truly a theory of these loops has been around for as long as such theories have been studied. Mandelstam, Poliakov, Wilson, among many others, have long argued that loop excitations should play a major role in quantum Yang-Mills theories, and that we must get to understand quantum Yang-Mills theories in terms of these excitations. In fact, much of the development of string theory has been influenced by this idea.

In *lattice* Yang-Mills theory, loop states have finite norm. In fact, certain finite linear combinations of loop states, called "spin network" states, form a well-defined and well-understood orthonormal basis in the Hilbert space of a lattice gauge theory.

However, in a QFT theory over a *continuous* background, the idea of formulating the theory in terms of loop-like excitations has never proved fruitful. The difficulty is essentially that loop states over a background are "too singular" and "too many". The quantum Maxwell state $|\alpha\rangle$ described above, for instance, has infinite norm; and an infinitesimal displacement of a loop state over the background spacetime produces a distinct, independent, loop state, yielding a continuum of loop states. Over a continuous background, the space spanned by the loop states is far "too big" for providing a basis of the (separable) Hilbert space of a QFT.

However, loop states are not too singular, nor too many, in a *background independent* theory. This is the key technical point on which LQG relies. The intuitive reason is the following. Spacetime itself is formed by loop-like states. Therefore the position of a loop state is relevant only *with respect to other loops*, and not with respect to the background. An infinitesimal (coordinate) displacement of a loop state does not produce a distinct quantum state, but only a gauge equivalent representation of the same physical state! Only a finite displacement carrying the loop state across another loop produces a physically different state. Therefore the size of the space of the loop states is dramatically reduced by diffeomorphism invariance: most of it is just gauge! Equivalently, we can think that a single loop has an intrinsic Planck size "thickness".

Therefore in a general relativistic context the loop basis becomes viable. The state space of the theory, called $\mathcal{K}_{\text{Diff}}$ is a separable Hilbert space spanned by loop states. More precisely, as we shall see in chapter 6, $\mathcal{K}_{\text{Diff}}$ admits an orthonormal basis of spin network states, which are formed by finite linear combinations of loop states, and are defined precisely as the spin network states of a lattice Yang Mills theory. This Hilbert space and the field operators that act on it are described in chapter 6. They form the basis of the mathematical structure of LQG.

Therefore LQG is the result of the convergence of two lines of thinking, each characteristic of XXth century theoretical physics. On the one hand, the intuition of Faraday, Yang and Mills, Wilson, Mandelstam, Poliakov... intuition that forces are described by lines. On the other hand, the Einstein-Wheeler-DeWitt intuition of background independence and background-independent quantum states. Truly remarkably, each of these two lines of thinking is the solution of the blocking difficulty of the other. On the one hand, the traditional non-viability of the loop basis in the continuum disappears because of background independence. On the other hand, the traditional difficulty of controlling diffeomorphism invariant quantities comes under control thanks to the loop basis.

Even more remarkably, the spin network states generated by this happy marriage turn out to have a surprisingly compelling geometric interpretation, which I sketch below.

1.2.2 Quantum space: spin networks

Physical systems reveal themselves by interacting with other systems. These interactions may happen in "quanta": energy is exchanged with an oscillator of frequency ν in discrete packets, or quanta, of size $E = h\nu$. If the oscillator is in the n-th energy eigenstate, we say that there are n quanta in it. If the oscillator is a mode of a free field, we say that there are n "particles" in the field. Therefore we can view the electromagnetic field as made by its quanta, the photons. What are the quanta of the gravitational field? Or, since the gravitational field is the same entity as spacetime, what are the quanta of space?

The properties of the quanta of a system are determined by the spectral properties of the operators representing the quantities involved in our interaction with the system. The operator associated to the energy of the oscillator, for instance, has discrete spectrum, and the number of quanta n labels its eigenvalues. The set of its eigenstates form a basis in the state space of the quantum system: this fact allows us to view each state of the system as a quantum superposition of states $|n\rangle$ formed by n quanta. To understand the quantum properties of space, we have therefore to consider the spectral problem of the operators associated to the quantities involved in our interaction with space itself. The most direct interaction we have with the gravitational field is via the geometrical structure of the physical space. A measurement of length, area, or volume is in fact, according to GR, a measurement of local properties of the gravitational field.

For instance, the volume $\mathbf{V}$ of a physical region R is

$$\mathbf{V} = \int_R d^3x \, |\det e(x)|, \tag{1.9}$$

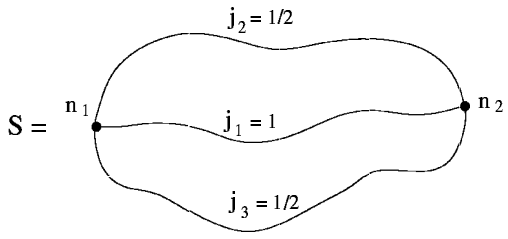


Figure 1.2: A simple spin network.

where $e(x)$ is the (triad matrix representing the) gravitational field. In quantum gravity, $e(x)$ is a field operator, and $\mathbf{V}$ is therefore an operator as well.

The volume $\mathbf{V}$ is a nonlinear function of the field e and the definition of the volume operator implies products of local operator valued distributions. This can be achieved as a limit, using an appropriate regularization procedure. The development of regularization procedures that remain meaningful in the absence of a background metric is a major technical tool on which LQG is based. Using these techniques, a well defined self adjoint operator $\mathbf{V}$ can be defined. The computation of its spectral properties is then one of the main results of LQG, and will be derived in section 6.6.5.

The spectrum of $\mathbf{V}$ turns out to be discrete. Therefore the spacetime volume manifest itself in quanta, of definite volume size, given by the eigenvalues of the volume operator. These quanta of space can be intuitively thought of as quantized "grains" of space. The first intuitive picture of quantum space is therefore that of "grains of space". These have quantized amounts of volume, determined by the spectrum of the operator $\mathbf{V}$.

The next element of the picture is the information on which grain is adjacent to which. Adjacency (being contiguous, being in touch, being nearby) is the basis of spatial relations. If two spacetime regions are adjacent, that is, if they touch each other, they are separated by a surface S. Let $\mathbf{A}$ be the area of the surface S. Area as well is a function of the gravitational field, and is therefore represented by an operator, like the volume. The spectral problem for this operator has been solved in LQG, as well. It is discussed in detail in Section 6.6.2. The spectrum turns out to be discrete as well. Intuitively, the grains of space are separated by "quanta of area". The principal series of the eigenvalues of the area, for instance, is labelled by multiplets of half integers $j_i, i = 1 \dots n$ and turns out to be given by

$$\mathbf{A} = 8\pi\gamma\, \hbar G \sum_i \sqrt{j_i(j_i+1)}. \tag{1.10}$$

where γ, the Immirzi parameter, is a free dimensionless constant of the theory.

Consider a quantum state of space $|s\rangle$ formed by N "grains" of space, some of which are adjacent to one another. Represent this state as an abstract graph Γ with N nodes: the nodes of the graph represent the grains of space; the links of the graph link adjacent grains and represent the surfaces separating two adjacent grains. The quantum state is then characterized by the graph Γ and by labels on nodes and on links: the label i_n on a node n is the quantum number of the volume and the label j_l on a link l is the quantum number of the area.

A graph with these labels is called an (abstract) "spin network" $s = (\Gamma, i_n, j_l)$. See Figure 1.2. In section 6.3.1, we will see that the quantum numbers i_n and j_l are determined by the representation theory of the local gauge group ($SU(2)$). More precisely, j_l labels unitary irreducible representations and i_n labels a basis in the space of the intertwiners between the representations adjacent to the node n. The area of a surface cutting n links of the spin network with labels $j_i, i = 1 \dots n$ is then given by (1.10).

As shown in section 6.3.1, the (kinematical) Hilbert space $\mathcal{K}_{\text{Diff}}$ admits a basis labeled precisely by these spin networks. This is a basis of states in which certain area and volume operators are

diagonal. Its physical interpretation is the one sketched above: a spin network state $|s\rangle$ describes a quantized three-geometry.

A loop state $|\alpha\rangle$ is a spin network state in which the graph Γ has no nodes, namely is a single loop α, and is labeled by the fundamental representation of the group. In such a state, the gravitational field has support just on the loop α itself, as the electric field in (1.6).

In LQG, physical space is a quantum superposition of spin networks, in the same sense in which the electromagnetic field is a quantum superposition of n-photon states. The first and basic prediction of the (free) QFT of the electromagnetic field is the existence of the photons, and the specific quantitative prediction of the energy and the momentum of the photons of a given frequency. Similarly, the first prediction of LQG is the existence of the quanta of area and volume, and the quantitative prediction of their spectrum.

The theory predicts that any sufficiently accurate measurement of area or volume would measure one of these spectral values. So far, verifying this prediction appears to be outside our technological capacities.

Where is a spin network? A spin network state does not have a position. It is an abstract graph – not a graph immersed in a spacetime manifold. Only abstract combinatorial relations defining the graph are significant, not its shape or its position in space.

In fact, a spin network state is not *in* space: it *is* space. It is not localized with respect to something else: something else (matter, particles, other fields) might be localized with respect to it. To ask "where is a spin network" is like asking "where is a solution of the Einstein equations". A solution of the Einstein equations is not "somewhere": it is the "where" with respect to which anything else can be localized. In the same way, the other dynamical objects, such as Yang-Mills and fermion fields, live on the spin network state.

This is a consequence of diffeomorphism invariance. Technically, spin network states are first defined as graphs embedded in a three dimensional manifold; then the implementation of the diffeomorphism gauge identifies two graphs that can be deformed into each other. They are gauge equivalent. This is like identifying two solutions of the Einstein equations that are related by a change of coordinates. Spin networks embedded in a manifold are denoted S and called "embedded spin networks"; equivalence classes of these under diffeomorphisms are called "abstract spin networks", or s-knots. A quantum state of space is determined by an s-knot.[7]

The fact that spin networks do not live *in* space, but rather *are* space, has long ranging consequences. Space itself turns out to have a discrete and combinatorial character. Notice that this is not imposed on the theory, or assumed. It is the result of a completely conventional quantum mechanical calculation of the spectrum of the physical quantities that describe the geometry of space. Since there is no spatial continuity at small scale, there is (literally!) no room in the theory for ultraviolet divergencies. The theory effectively cuts itself off at the Planck scale. Space is effectively granular at the Planck scale, and there is no ultraviolet limit.

Chapter 7 describes how Yang-Mills and fermion fields can be coupled to the theory. This can be obtained by enriching the structure of the spin networks s. In the case of a Yang-Mills theory with gauge group G, for instance, links carry an additional quantum number, labeling irreducible representations G. The spin network itself behaves like the lattice of lattice Yang-Mills theory. In quantum gravity, therefore, the lattice itself becomes a dynamical variable. But notice a crucial difference with respect to conventional lattice Yang-Mills theory: the lattice size is not to be scaled down to zero: it has physical Planck size.

[7] The expression "spin network" is used in the literature to designate both the embedded and the abstract ones, as well as to designate the quantum states they label.

In summary, spin networks provide a mathematically well defined and physically compelling description of the kinematics of the quantum gravitational field. They also provide a well-defined picture of the small scale structure of space. It is remarkable that this novel picture of space emerges simply from the combination of old Yang-Mills theory ideas with general relativistic background independence.

1.2.3 Dynamics in background independent QFT

The dynamics of the quantum gravitational field can be described giving amplitudes $W(s)$ for spin network states. Let me illustrate here, in a heuristic manner, the physical interpretation of these amplitudes and the way they are defined in the theory.

Interpretation of the amplitude $W(s)$. The quantum dynamics of a particle is entirely described by the transition probability amplitudes

$$W(x,t,x',t') = \langle x|e^{-\frac{i}{\hbar}H_0(t-t')}|x'\rangle = \langle x,t|x',t'\rangle, \tag{1.11}$$

where $|x,t\rangle$ is the eigenstate of the Heisenberg position operator $x(t)$ with eigenvalue x, H_0 is the hamiltonian operator and $|x\rangle = |x,0\rangle$. The propagator $W(x,t;x',t')$ depends on two events (x,t) and (x',t') that bound a finite portion of a classical trajectory. The space of the pairs of events (x,t,x',t') is called $\mathcal{G}$ in this book.

A physical experiment consists in a preparation at time t' and a measurement at time t. Say that in a particular run we have localized the particle in x' at t' and then found it in x at time t. The set (x,t,x',t') represents the complete set of data of a specific complete observational set up, including preparation and measurement. The space $\mathcal{G}$ is the space of these data sets. In the quantum theory, we associate associate the complex amplitude $W(x,t,x',t')$, which is a function on $\mathcal{G}$, to any such data set. As emphasized by Feynman, this amplitude codes the full quantum dynamics. Following Feynman, we can compute $W(x,t,x',t')$ with a sum-over-paths that take the values x and x' at t and t'.

If we measure a different observable than position, we obtain states different from the states $|x\rangle$. Let $|\psi_{\text{in}}\rangle$ be the state prepared at time t', and let $|\psi_{\text{out}}\rangle$ be the state measured at time t. The amplitude associated to these measurements is

$$A \;=\; \langle\psi_{\text{out}}|e^{-\frac{i}{\hbar}H_0(t-t')}|\psi_{\text{in}}\rangle. \tag{1.12}$$

The pair of states $(\psi_{\text{in}},\psi_{\text{out}})$ determines a state $\psi = |\psi_{\text{in}}\rangle \otimes \langle\psi_{\text{out}}|$ in the space $\mathcal{K}_{t,t'}$ which is the tensor product of the Hilbert space of the initial states and (the dual of) the Hilbert space of the final states. The propagator defines a (possibly generalized) state $|0\rangle$ in $\mathcal{K}_{t,t'}$, by $\langle 0|\Big(|x'\rangle \otimes \langle x|\Big) = W(x,t,x',t')$. The amplitude (1.12) can be written simply as

$$A \;=\; \langle 0|\psi\rangle\ . \tag{1.13}$$

Therefore we can express the dynamics from t' to t in terms of a single state $|0\rangle$ in a Hilbert space $\mathcal{K}_{t,t'}$ that represents outcomes of measurements on t' *and* t. The state $|0\rangle$ is called the covariant vacuum, and should not be confused with the state of minimal energy.

Let us extend this idea to field theory. In field theory, the analog of the data set (x,t,x',t'), is the couple $[\Sigma,\varphi]$, where Σ is a 3d surface Σ bounding a finite spacetime region, and φ is a field configuration on Σ. These data define a set of events $(x\in\Sigma,\varphi(x))$ that bound a finite portion of a classical configuration of the field, like (x,t,x',t') bound a finite portion of the classical trajectory of the particle. The data of a local experiment (measurements, preparation, or just assumptions) must

in fact refer to the state of the system on the entire boundary of a finite spacetime region. The field theoretical space $\mathcal{G}$ is therefore the space of surfaces Σ and field configurations φ on Σ. Quantum dynamics can be expressed in terms of an amplitude $W[\Sigma, \varphi]$. Following Feynman's intuition, we can formally define $W[\Sigma, \varphi]$ in terms of a sum over field configurations that take the value φ on Σ. In fact, in section 5.3, I argue that the functional $W[\Sigma, \varphi]$ captures the dynamics of a QFT.

Notice that the dependence of $W[\Sigma, \varphi]$ on the geometry of Σ codes the spacetime position of the measuring apparatus. In fact, the relative position of the components of the apparatus is determined by their physical distance and the physical time lapsed between measurements, and these data are contained in the metric of Σ.

Consider now a background independent theory. Diffeomorphism invariance implies immediately that $W[\Sigma, \varphi]$ is independent from Σ. This is the analog of the independence of $W(x, y)$ from x and y, mentioned above in section 1.1.4. Therefore in gravity W depends only on the boundary value of the fields. However, the fields include the gravitational field, and the gravitational field determines the spacetime geometry. Therefore the dependence of W on the fields is still sufficient to code the relative distance and time separation of the components of the measuring apparatus!

What is happening is that in background dependent QFT we have two kinds of measurements: the ones that determine the distances of the parts of the apparatus and the time lapsed between measurements, and the actual measurements of the fields' dynamical variables. In quantum gravity, instead, distances and time separations are on the same ground as the dynamical fields. This is the core of the general relativistic revolution, and the key for background independent QFT.

We need one final step. Notice from (1.11) that the argument of W is not the classical quantity, but rather the eigenstate of the corresponding operator. The eigenstates of the gravitational field are spin networks. Therefore in quantum gravity the argument of W must be a spin network, representing the possible outcome of a measurement of the gravitational field (or the geometry) on a closed 3d surface. Therefore in quantum gravity physical amplitudes must be expressed by amplitudes of the form $W(s)$. These give the correlation probability amplitude associated to the outcome s in a measurement of a geometry, as $W(x, t, x', t')$ does for a particle.

A particularly interesting case is when we can separate the boundary surface in two components, then $s = s_{\text{out}} \cup s_{\text{in}}$. In this case, $W(s_{\text{out}}, s_{\text{in}})$ can be interpreted as the probability amplitude of measuring the quantum three-geometry s_{out} if s_{in} was observed.

Notice that a spin network s_{in} is the analog of (x, t), not just x alone. The independent time variable is mixed up with the physical variables (Chapter 3.) The notion of unitary quantum evolution in time is ill defined in this context, but probability amplitudes remain well defined and physically meaningful (Chapter 5.) The quantum dynamical information of the theory is entirely contained in the spin network amplitudes $W(s)$. Given a configuration of space and matter, these amplitudes determine a correlation probability of observing it.

Calculation of the amplitude $W(s)$. In the relativistic formulation of classical hamiltonian theory, dynamics is governed by the relativistic hamiltonian H.[8] This is discussed in detail in Chapter 3. The quantum dynamics is governed by the corresponding quantum operator H. In quantum gravity, H is defined on the space of the spin networks. There is no external time variable t in the theory, and the quantum dynamical equation which replaces the Schrödinger equation is the equation $H\Psi = 0$, called the Wheeler-DeWitt equation. The space of the solutions of the Wheeler-DeWitt equation is denoted $\mathcal{H}$. There is an operator $P : \mathcal{K}_{\text{Diff}} \to \mathcal{H}$ that projects $\mathcal{K}_{\text{Diff}}$ on the solutions of the Wheeler-DeWitt equation (for a mathematically more precise statement, see section 5.2).

[8] H is sometimes called the "hamiltonian constraint" or the "superhamiltonian".

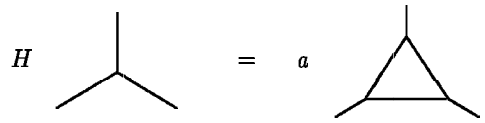


Figure 1.3: Scheme of the action of H on a node of a spin network.

The transition amplitudes $W(s, s')$ are the matrix elements of the operator P. They define the physical scalar product, namely the scalar product of the space $\mathcal{H}$

$$W(s, s') = \langle s|P|s'\rangle_{\mathcal{K}_{\text{Diff}}} = \langle s|s'\rangle_{\mathcal{H}}. \tag{1.14}$$

Thus, the transition amplitude between two states is simply their physical scalar product (Chapter 5.) More in general, there is a preferred state $|\emptyset\rangle$ in $\mathcal{K}_{\text{Diff}}$ which is formed by no spin networks. It represents a space with zero volume, or, more precisely, no space at all. The covariant vacuum state which defines the dynamics of the theory is defined by $|0\rangle = P|\emptyset\rangle$. The amplitude of a spin network is defined by

$$W(s) \;=\; \langle 0|s\rangle \;=\; \langle \emptyset|P|s\rangle \;. \tag{1.15}$$

The construction of the operator H is a major task in LQG. It is delicate and it requires a nontrivial regularization procedure in order to deal with operator products. Chapter 7 is devoted to this construction. Remarkably, the limit in which the regularization is removed exists precisely thanks to diffeomorphism invariance (Section 7.1.) This is a second major payoff of background independence. At present, more than one version of the operator H has been constructed, and it is not yet clear which variant (if any!) is correct. The remarks that follow refers to all of them.

The most remarkable aspect of the Hamiltonian operator H is that it acts only on the nodes. A state labeled by a spin network without nodes –that is, in which the graph Γ is simply a collection of non intersecting loops– is a solution of the Wheeler-DeWitt equation. In fact, the unexpected fact that exact solutions of the Wheeler-DeWitt equation could be found at all was the first major surprise that raised interest in LQG in the first place, in the late eighties.

Acting on a generic state $|s\rangle$, the action of the operator H turns out to be discrete and combinatorial: the topology of the graph is changed and the labels are modified, in the vicinity of a node. A typical example of the action of H on a node is illustrated in Figure 1.3: the action on a node splits the node in three nodes and multiplies the state by a number a (that depends on the labels of the spin network around the node). Labels of links and nodes are not indicated in the figure.

Notice the various manners in which the spin network basis is effective in quantum gravity. The states in the spin network basis

i. diagonalize area and volume.

ii. control diff-invariance: diffeomorphism equivalence classes of state are labelled by the s-knots.

iii. simplify the action of H, reducing it to a combinatorial action on the nodes.

The construction of the hamiltonian operator H completes the definition of the general formalism of LQG in the case of pure gravity. This is extended to matter couplings in Chapter 7. In Chapter 8, I describe some of the most interesting applications of the theory: in particular, the application to the initial cosmological singularity, to derive the thermodynamical behavior of black holes, and to compute possible quantum gravity effects on particle propagation.

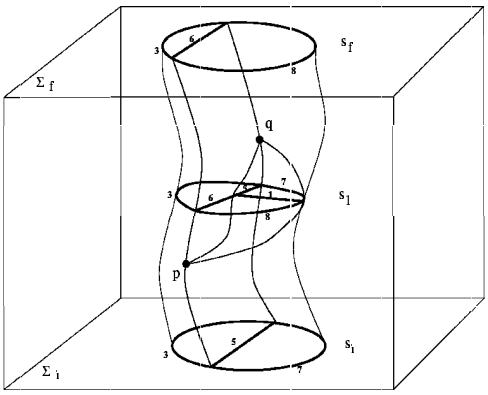


Figure 1.4: A spinfoam.

1.2.4 Quantum spacetime: spinfoam

To be able to compute anything we want from a theory, it is not sufficient to have the general definition of a theory. A road towards the calculation of generic transition amplitudes in quantum gravity is provided by the spinfoam formalism.

Following Feynman's ideas, we can give $W(s, s')$ a representation as a sum over paths. This representation can be obtained in various manners. In particular, it can be intuitively derived from a perturbative expansion, summing over different histories of sequences of actions of H that send s' into s.

A path is then the worldhistory of a graph, with interactions happening at the nodes. This worldhistory is a two-complex, as in Figure 1.4 namely a collection of faces (the world-histories of the links); faces join at edges (the world histories of the nodes); in turn, edges join at vertices. A vertex represents an individual action of H. An example of vertex, corresponding to the action of H of Figure 1.3, is illustrated in Figure 1.5. Notice that moving from the bottom to the top, a section of the two-complex goes precisely from the graph on the left hand side of Figure 1.3 to the one on the right hand side. Thus, a two-complex is like a Feynman graph, but with one additional structure. A Feynman graph is composed by vertices and edges, a spinfoam by vertices, edges and faces.

Faces are labelled by the area quantum numbers j_l and edges by the volume quantum numbers i_n. A two-complex with faces and edges labelled in this manner is called a "spinfoam" and denoted σ. Thus, a spinfoam is a Feynman graph of spin networks, or a world history of spin networks. A history going from s' to s is a spinfoam σ bounded by s' and s.

In the perturbative expansion of $W(s, s')$, there is a term associated to each spinfoam σ bounded by s and s'. This term is called the amplitude of σ. The amplitude of a spinfoam turns out to be given by (a measure term $\mu(\sigma)$ times) the product over the vertices v of a vertex amplitude $A_v(\sigma)$. The vertex amplitude is determined by the matrix element of H between the incoming and the outgoing spin networks and is a function of the labels of the faces and the edges adjacent to the vertex. This is analogous to the amplitude of a conventional Feynman vertex, which is determined

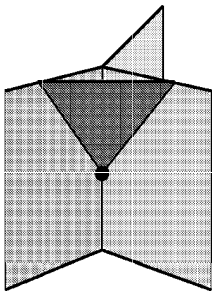


Figure 1.5: The vertex of a spinfoam.

by the matrix element of the Hamiltonian between the incoming and outgoing states.

The physical transition amplitudes $W(s,s')$ are then obtained by summing over spinfoams bounded by the spin networks s and s'.

$$W(s,s') \sim \sum_{\substack{\sigma \\ \partial\sigma = s \cup s'}} \mu(\sigma) \prod_v A_v(\sigma). \tag{1.16}$$

More generally for a spin network s representing a closed surface

$$W(s) \sim \sum_{\substack{\sigma \\ \partial\sigma = s}} \mu(\sigma) \prod_v A_v(\sigma). \tag{1.17}$$

In general, the Feynman path integral can be derived from Schrödinger theory by exponentiating the Hamiltonian operator, but it can also be directly interpreted as a sum over classical trajectories of the particle. Similarly, the spinfoam sum (1.16) can be interpreted as a sum over spacetimes. That is, the sum (1.16) can be seen as a concrete and mathematically well defined realization of the (ill defined) Wheeler-Misner-Hawking representation of quantum gravity as a sum over four-geometries

$$W({}^3g, {}^3g') \sim \int_{\partial g = {}^3g \cup {}^3g'} [Dg]\ \mu(g)\ e^{\frac{i}{\hbar} S[g]}. \tag{1.18}$$

Because of their foamy structure at the Planck scale, spinfoams can be viewed as a mathematically precise realization of Wheeler's intuition of a spacetime "foam". In chapter 9, I describe various concrete realizations of equation (1.16), as well as the possibility to directly relate (1.16) with a discretization of (1.18).

1.3 Conceptual issues

The search for a quantum theory of gravity raises questions such as: What is space? What is time? What is the meaning of "being somewhere"? What is the meaning of "moving"? Is motion to be defined with respect to objects or with respect to space? Can we formulate physics without referring to time or to spacetime? And also: What is matter? What is causality? What is the role of the observer in physics?

Questions of this kind have played a central role in periods of major advances in physics. For instance, they played a central role for Einstein, Heisenberg, Bohr and their colleagues. But also for Descartes, Galileo, Newton and their contemporaries, and for Faraday, Maxwell and their colleagues. Today, this manner of posing problems is often regarded as "too philosophical" by many physicists.

Most physicists of the second half of the XXth century, indeed, have viewed questions of this nature as irrelevant. This view was appropriate for the problems they were facing: one does not need to worry about first principles in order to apply the Schrödinger equation to the helium atom, to understand how a neutron star stays together, or to find out the symmetry group governing the strong interactions. In this period physicists lost interest in general issues. As was said during this period, "do not ask what the theory can do for you; ask what you can do for the theory". That is, do not ask foundational questions, just keep developing and adjusting the theory you happen to find in front of you. When the basics are clear and the issue is problem-solving within a given conceptual scheme, there is no reason to worry about foundations: the problems are technical and the pragmatical approach is the most effective one.

Today the kind of difficulties that we face has changed. To understand quantum spacetime, we have to return, once more, to those foundational issues. We have to find new answers to the old foundational questions. The new answers have to take into account what we have learned with QM and GR. This conceptual approach is not the one of Weinberg and Gell-Mann, but it is the one of Newton, Maxwell, Einstein, Bohr, Heisenberg, Faraday, Boltzmann and so many others. It is clear from the writings of these scientists that they have discovered what they have discovered in thinking about general foundational questions. The problem of quantum gravity will not be solved unless we reconsider these questions.

Several of these questions are discussed in the text. Here I only comment on one of these conceptual issues: the role of the notion of time.

1.3.1 Physics without time

The transition amplitudes $W(s, s')$ do not depend explicitly on time. This is to be expected, because the physical predictions of classical general relativity do not depend explicitly on the time coordinate t either. The theory predicts correlations between physical variables, not the way physical variables evolve with respect to a preferred time variable. But what is the meaning of a physical theory in which the time variable t does not appear?

Let me tell a story. It was Galileo Galilei who first realized that physical motion of objects on Earth could be described by mathematical laws expressing the evolution of observable quantities A, B, $C\ldots$ in time. That is, laws for the functions $A(t)$, $B(t)$, $C(t)\ldots$ A crucial contribution by Galileo was to find an effective way to measure the time variable t, and therefore provide an operational meaning to these functions. In fact, Galileo gave a decisive contribution to the discovery of the modern clock, by realizing, as a young man, that the small oscillations of a pendulum "take equal time". The story goes that Galileo was staring at the slow oscillations of the big chandelier that can still be seen in the marvelous Cathedral of Pisa.[9] He checked the period of the oscillations against his pulse and realized that the same number of pulses lapsed during any oscillation of the pendulum. This was the key insight at the basis of the modern clock: today virtually every clock contains an oscillator. Later in life, Galileo used a pendulum as a clock to discover the first quantitative terrestrial physical law in his historic experiments on the fall along inclines.

Now, the puzzling part of the story is that while Galileo checked the pendulum against his pulse, not long later doctors were checking their patient's pulse against a pendulum. What is the actual meaning of the pendulum periods taking "equal time"? An equal amount of t lapses in any oscillation: how do we know this, if we can access t only via another pendulum?

It was Newton who cleared up the issue conceptually. Newton *assumes* that a *unobservable* quantity t exists, which flows ("absolute and equal to itself"). We write equations of motion in terms of this t, but we cannot truly access t: we can build clocks that give readings $T_1(t)$, $T_2(t)$,

[9] Nice story. Too bad the chandelier was hung there a few decades after Galileo's discovery.

...that, according to our equations, approximate t with the precision we want. What we actually measure is the evolution of other variables against clocks, namely $A(T_1)$, $B(T_1)$. Furthermore, we can check clocks against one another by measuring the functions $T_1(T_2)$, $T_2(T_3)$... The fact that all these observations agree with what we compute using evolution equations in t give us confidence in the method. In particular, it give us confidence that to assume the existence of the *unobservable* physical quantity t is a useful and reasonable thing to do.

Simply: the usefulness of this assumption is lost in quantum gravity. The theory allows us to calculate the relations between observable quantities, such as $A(B)$, $B(C)$, $A(T_1)$, $T_1(A)$, ..., which is what we see. But it does not give us the evolution of these observable quantities in terms of an unobservable t, as Newton theory and special relativity do. In a sense, this simply means that there are no good clocks at the Planck scale.

Of course, in a specific problem we can choose one variable, decide to treat it as the independent variable, and call it "the" time. For instance a certain clock time, a certain proper time along a certain particle history, or else. The choice is largely arbitrary and generally it is only locally meaningful. A generally covariant theory does not choose a preferred time variable.

Here are two examples to illustrate this arbitrariness.

- Imagine we throw a precise clock upward and compare its reading t_f when it lands back, with the reading t_e of a clock on earth. GR predicts that the two clocks read differently, and provides a quantitative relation between t_f and t_e. Is this about the observable t_f evolving in the physical time t_e, or about the observable t_e evolving in the physical time t_f?[10]

- The cosmological context is often indicated as one in which a natural choice of time is available: the cosmological time t_c is the proper time from the big bang along the galaxies' world lines. But an event A happening on Andromeda at the same t_c as ours happens *much later* than an event B on Andromeda simultaneous to us in the sense of Einstein's definition of simultaneity.[11] So, what is happening "right now" on Andromeda? A or B? Furthermore, the real world is not trully homogeneous: when two galaxies, having two different ages from the big-bang merge, which of the two has the right time?

As far as we remain within *classical* general relativity, a given gravitational field has the structure of a pseudo-Riemannian manifold. Therefore, the dynamics of the theory has no preferred time variable, but we nevertheless have a notion of spacetime for each given solution. But in quantum theory there are no classical field configurations, like there are no trajectories of a particle. Thus, in quantum gravity the notion of spacetime disappears in the same manner in which the notion of trajectory disappears in the quantum theory of a particle. A single spinfoam can be thought of as representing a spacetime, but the history of the world is not a single spinfoam: it is a sum over spinfoams.

The theory is conceptually well-defined without making use of the notion of time. It provides probabilistic predictions for correlations between the physical quantities that we can observe. In principle we can check these predictions against experiments. Furthermore, the theory provides a clear and intelligible picture of the quantum gravitational field, namely of a "quantum geometry".

Thus, there is no background "spacetime", forming the stage on which things move. There is no "time" along which everything flows. The world in which we happen to live can be understood without using the notion of time.

---

[10] If you are tempted to say that the clock t_e on Earth gives the "true time" recall that the Riemannian distance between the two events at which the clocks meet is t_f, not t_e: it is the clock going up and down that follows a geodesic.

[11] Thanks to Marc Lachieze-Rey for this observation.

Bibliographical note

The fact that perturbative quantum general relativity is nonrenormalizable has been long believed, but was proven only in 1986 by Goroff and Sagnotti in [27].

For an orientation on current research on quantum gravity, see for instance the review papers [28, 29, 30, 31]. An interesting panoramic of points of view on the problem is in the various contributions to the book [32]. I have given a critical discussion on the present state of spacetime physics in [33, 34, 35]. A historical account of the development of the quantum gravity is given below in appendix B.

As a general introduction to quantum gravity –a subject where nothing yet is certain– the student eager to learn is strongly advised to study also the old classic reviews, which are rich in ideas and present different points of view, such as John Wheeler 1967 [36], Steven Weinberg 1979 [37], Stephen Hawking 1979 and 1980 [38, 39], Karel Kuchar 1980 [40], and Chris Isham's magistral syntheses [41, 42, 43]. On string theory, classic textbooks are Green, Schwarz and Witten, and Polchinksi [44]. For a discussion of the difficulties of string theory and a comparison of the results of strings and loops, see [45], written in the form of a dialog, and [46]. For a fascinating presentation of Alain Connes' vision, see [47].

Lee Smolin's popularization book [48] provides a readable and enjoyable introduction to LQG.

LQG has inspired novels and short stories. *Blue Mars*, by Kim Stanley Robinson [49], contains a description of the future evolution and merge of loop gravity and strings. I recommend the science fiction novel *Schild Ladder*, by Greg Egan [50], which opens with one of the most clear presentations of the picture of space given by loop gravity (Greg is a talented writer and also a scientists who is contributing to the development of LQG), and, for those who can read Italian, *Anna prende il volo*, by Enrico Palandri [51], a charming novel with a gentle meditation on the meaning of the disappearance of time. Literature has the capacity of delicately merging the novel hard views that science develops into the common discourse of our civilization.

Chapter 2

General Relativity

Lev Landau has called GR "the most beautiful" of the scientific theories. The theory is first of all a description of the gravitational force. Nowadays it is very extensively supported by terrestrial and astronomical observations, and so far it has never been questioned by an empirical observation.

But GR is far more than that. It is a complete modification of our understating of the basic grammar of nature. This modification does not regard the sole gravitational interaction: it regards all aspects of physics. In fact, the extent to which Einstein's discovery of this theory has modified our understanding of the physical world and the full reach of its consequences have not been completely unraveled yet.

This chapter is not an introduction to GR, nor an exhaustive description of the theory. For this I refer the reader to the classic textbooks on the subject. Here, I give a short presentation of the formalism in a compact and modern form, emphasizing the reading of the theory which is most useful for quantum gravity. I also discuss in detail the physical and conceptual basis of the theory, and the way it has modified our understanding of the physical world.

2.1 Formalism

2.1.1 Gravitational field

Let M be the "spacetime" four-dimensional manifold. Coordinates on M are written as $x, x', \ldots$, where $x = (x^\mu) = (x^0, x^1, x^2, x^3)$. Indices $\mu, \nu, \ldots = 0, 1, 2, 3$ are spacetime tangent indices.

The gravitational field e is a one-form

$$e^I(x) = e^I_\mu(x)\, dx^\mu \tag{2.1}$$

with values in Minkowski space. Indices $I, J, \ldots = 0, 1, 2, 3$ label the components of a Minkowski vector. They are raised and lowered with the Minkowski metric η_{IJ}. The reason that led Einstein to understand that the gravitational field has this form are discussed in section 2.2.3.

I call "gravitational field" the tetrad field rather than Einstein's metric field $g_{\mu\nu}(x)$. There are three reasons for this: (i) the standard model cannot be written in terms of g because fermions require the tetrad formalism; (ii) the tetrad field e is nowadays more utilized than g in quantum gravity; and (iii) I think that e represents the gravitational field in a more conceptually clean way than g (see section 2.2.3.) The relation with the metric formalism is given in section 2.1.5.

The spin connection ω is a one-form with values in the Lie algebra of the Lorentz group $so(3,1)$

$$\omega^I{}_J(x) = \omega^I_{\mu J}(x)\, dx^\mu, \tag{2.2}$$

where $\omega^{IJ} = -\omega^{JI}$. It defines a covariant partial derivative D_μ on all fields that have Lorentz (I) indices:

$$D_\mu v^I = \partial_\mu v^I + \omega^I_{\mu J}\, v^J \tag{2.3}$$

and a gauge covariant exterior derivative D on forms. For instance, for a one-form u^I with a Lorentz index,

$$Du^I = du^I + \omega^I{}_J \wedge u^J. \tag{2.4}$$

The torsion two-form is defined as

$$T^I = De^I = de^I + \omega^I{}_J \wedge e^J. \tag{2.5}$$

A tetrad field e determines uniquely a torsion free spin connection $\omega = \omega[e]$, called compatible with e, by

$$T^I = de^I + \omega[e]^I{}_J \wedge e^J = 0. \tag{2.6}$$

The explicit solution of this equation is given below in (2.88) or (2.89).

The curvature R of ω is the Lorentz algebra valued two-form[1]

$$R^I{}_J = R^I{}_{J\,\mu\nu}\; dx^\mu \wedge dx^\nu \tag{2.7}$$

defined by[2]

$$R^I{}_J = d\omega^I{}_J + \omega^I{}_K \wedge \omega^K{}_J. \tag{2.8}$$

A region where the curvature is zero is called "flat". Equations (2.5) and (2.8) are called the Cartan structure equations.

The Einstein equations "in vacuum" are

$$\epsilon_{IJKL}\; (e^I \wedge R^{JK} + \lambda\; e^I \wedge e^J \wedge e^K) = 0. \tag{2.9}$$

The equation (2.6) relating e and ω and the Einstein equations (2.9) are the field equations of GR in the absence of other fields. They are the Euler-Lagrange equations of the action

$$S[e,\omega] = \frac{1}{16\pi G} \int \epsilon_{IJKL}\;\; (e^I \wedge e^J \wedge R[\omega]^{KL} + \lambda\; e^I \wedge e^J \wedge e^K \wedge e^L). \tag{2.10}$$

G is the Newton constant[3]; λ is the cosmological constant, which I often set to zero below.

Inverse tetrad. Using the matrix $e^\mu_I(x)$, defined as the inverse of the matrix $e^I_\mu(x)$, we define the Ricci tensor

$$R^I_\mu \;\; = \;\; R^{IJ}{}_{\mu\nu}\; e^\nu_J, \tag{2.11}$$

and the Ricci scalar

$$R \;\; = \;\; R^I_\mu\; e^\mu_I, \tag{2.12}$$

and write the vacuum Einstein equations (2.9) as

$$R^I_\mu - \frac{1}{2}\,(R+\lambda) e^I_\mu = 0. \tag{2.13}$$

---

[1] Generally I write spacetime indices $\mu\nu$ before internal Lorentz indices IJ. But for the curvature I prefer to stay closer to Riemann's notation.

[2] Sometimes the curvature of a connection $\omega^I{}_J$ is written as $R^I{}_J = D\omega^I{}_J$. If we naively use the definition (2.3) for D, we get an extra 2 in the quadratic term. The point is that the indices on the connection are not vector indices. That is, (2.3) defines the action of D on sections of a vector bundle, and a connection is not a section of a vector bundle.

[3] The constant $16\pi G$ has no effect on the classical equations of motion (2.9). However, it governs the strength of the interaction with the matter fields described below, and it also determines the quantum properties of the system. In this it is similar to the mass constant m in front of a free particle action: the classical equations of motion ($\ddot{x} = 0$) do not depend on m, but the quantum dynamics of the particle does. For instance, the rate at which a wave packet spreads depends on m. Similarly, we will see that the quanta of pure gravity are governed by this constant.

Second order formalism. Replacing ω with $\omega[e]$ in (2.10) we get the equivalent action

$$S[e] = \frac{1}{16\pi G}\int \epsilon_{IJKL}\ \ (e^I \wedge e^J \wedge R[\omega[e]]^{KL} + \lambda\, e^I \wedge e^J \wedge e^K \wedge e^L). \tag{2.14}$$

The formalism in (2.10) where e and ω are independent is called the first order formalism. The two formalism are not equivalent in the presence of fermions; we do not know which one is physically correct, because the effect of gravity on single fermions is hard to measure.

Selfdual formalism. Consider the selfdual "projector" P^i_{IJ}

$$P^i_{jk} = \frac{1}{2}\epsilon^i_{jk}, \qquad P^i_{0j} = -P^i_{j0} = \frac{i}{2}\delta^i_j. \tag{2.15}$$

where $i = 1, 2, 3$.[4] Define the the complex $SO(3)$ connection

$$A^i_\mu = P^i_{IJ}\ \omega^{IJ}_\mu. \tag{2.16}$$

Equivalently,

$$A^i = \omega^i + i\omega^{0i}. \tag{2.17}$$

We can use the complex selfdual connection A^i (3 complex one-forms), instead of the real connection $\omega^I{}_J$ (6 real one-forms), as the dynamical variable for GR. (This is equivalent to describing a system with two real degrees of freedom x and y in terms of a single complex variable $z = x + iy$.) In terms of A^i, the Einstein equations read

$$P_{iIJ}\ e^I \wedge (F^i + \lambda\ P^i_{KL} e^K \wedge e^L) = 0. \tag{2.18}$$

where $F^i = dA^i + \epsilon^i_{jk} A^j A^k$ is the curvature of A.[5] These are the Euler-Lagrange equations of the action

$$S[e, A] = \frac{1}{16\pi G}\int (iP_{iIJ}\ \ e^I \wedge e^J \wedge F^i + \lambda\ \epsilon_{IJKL} e^I \wedge e^J \wedge e^K \wedge e^L), \tag{2.19}$$

which can be obtained adding to the action (2.10) an imaginary term that does not change the equations of motion. The selfdual formalism is often used in canonical quantization, because it simplifies the form of the hamiltonian theory.

*** Plebanski formalism.** The Plebanski selfdual two-form is defined as

$$\Sigma^i = P^i_{IJ}\ e^I \wedge e^J. \tag{2.20}$$

That is

$$\Sigma^1 = e^2 \wedge e^3 + i e^0 \wedge e^1. \tag{2.21}$$

and so on cyclically. A straightforward calculation shows that Σ satisfies

$$D\Sigma^i \equiv d\,\Sigma^i + A^i{}_j \wedge \Sigma^j = 0. \tag{2.22}$$

The algebraic equations for a triplet a of complex two-forms Σ^i

$$3\ \Sigma^i \wedge \Sigma^j = \delta^{ij}\ \Sigma_k \wedge \Sigma^k = \delta^{ij}\ \overline{\Sigma}_k \wedge \overline{\Sigma}^k, \qquad \Sigma^i \wedge \overline{\Sigma}^j = 0 \tag{2.23}$$

are solved by equation (2.20), where e^I is an arbitrary real tetrad. The GR action can thus be written as

$$S[\Sigma, A] = \frac{i}{16\pi G}\int \left(\Sigma_i \wedge F^i + \lambda\ \Sigma_k \wedge \Sigma^k\right), \tag{2.24}$$

where Σ^i satisfies the Plebanski constraints (2.23). The Plebanski formalism is often used as starting point for spinfoam models.

2.1.2 "Matter"

In the general relativistic parlance, "matter" is anything else which is not the gravitational field. As far as we know, the world is made by the gravitational field, Yang-Mills fields, fermion fields and, presumably, scalar fields.

---

[4] The complex Lorentz algebra splits into two complex $so(3)$ algebras, called the selfdual and anti-selfdual components: $so(3,1,C) = so(3,C) \oplus so(3,C)$. The projector (2.15) reads out the selfdual component.

[5] Because of the split mentioned in the previous footnote, the curvature of the selfdual component of the connection is the selfdual component of the curvature.

Maxwell. The electromagnetic field is described by the one-form field A, the Maxwell potential

$$A(x) = A_\mu(x)\ dx^\mu \tag{2.25}$$

Its curvature is the two-form $F = dA$, with components $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. Its dynamics is governed by the action

$$S_{\rm M}[e, A] = \frac{1}{4}\int F^* \wedge F. \tag{2.26}$$

Yang-Mills. The above generalizes to a non-abelian connection A in a Yang-Mills group G. A defines a gauge covariant exterior derivative D and curvature F. The action is

$$S_{\rm YM}[e, A] = \frac{1}{4}\int Tr[F^* \wedge F] \tag{2.27}$$

where Tr is a trace on the algebra.

Scalar fields. Let $\varphi(x)$ be a scalar field, possibly with values in a representation of G. The Yang-Mills field A defines the covariant partial derivative

$$D_\mu\varphi = \partial_\mu\varphi + A^A_\mu L_A\varphi, \tag{2.28}$$

where L_A are the generators of the gauge algebra in the representations to which φ belongs. The action that governs the dynamics of the field is

$$S_{\rm sc}[e, A, \varphi] = \int d^4x\ e\left(\eta^{IJ}\ e^\mu_I\overline{D_\mu\varphi}\ e^\nu_J D_\nu\varphi + V(\varphi)\right). \tag{2.29}$$

where e is the determinant of e^I_μ and $V(\varphi)$ is a self interaction potential.

Fermions. A fermion field ψ is a field in a spinor representation of the Lorentz group, possibly with values in a representation of G. The spin connection ω and the Yang-Mills field A define the covariant partial derivative

$$D_\mu\psi = \partial_\mu\psi + \omega^I_{\mu J}L^J_I\psi + A^A_\mu L_A\psi, \tag{2.30}$$

where L^J_I and L_A are the generators of the Lorentz and gauge algebras in the representations to which ψ belongs. Define

$$\not{D}\psi = \gamma^I e^\mu_I D_\mu\psi \tag{2.31}$$

where γ^I are the standard Dirac matrices. The action that governs the dynamics of the fermion field is

$$S_{\rm f}[e, \omega, A, \varphi, \psi] = \int d^4x\ e\left(\bar\psi\not{D}\psi + Y(\varphi, \bar\psi, \psi)\right). \tag{2.32}$$

where the second term is a polynomial interaction potential with a scalar field.

The "lagrangian of the world": the standard model. As far as we know, the world can be described in terms of a set of fields $e, \omega, A, \psi, \varphi$, where $G = SU(3) \times SU(2) \times U(1)$, and ψ and φ are in suitable multiplets, and is governed by the action

$$\begin{aligned} S[e, \omega, A, \psi, \varphi] &= S_{\rm RG}[e, \omega] + S_{\rm YM}[e, A] + S_{\rm f}[e, \omega, A, \psi] + S_{\rm sc}[e, A, \psi, \varphi] \\ &= S_{\rm RG}[e, \omega] + S_{\rm matter}[e, \omega, A, \varphi, \psi], \end{aligned} \tag{2.33}$$

with suitable polynomials V and Y. The equations of motion that follow from this action by varying e are the Einstein equations (2.9) with a source term, namely

$$\epsilon_{IJKL}\ (e^I \wedge R^{JK} + \lambda\ e^I \wedge e^J \wedge e^K) = 8\pi G\ T_I. \tag{2.34}$$

where the energy-momentum three-form

$$T_I = T_I^\mu\ \epsilon_{\mu\nu\rho\sigma} dx^\nu \wedge dx^\rho \wedge dx^\sigma \tag{2.35}$$

is defined by

$$T_I^\mu(x) = \frac{\delta S_{\text{matter}}}{\delta e^I_\mu(x)}. \tag{2.36}$$

Equivalently, the Einstein equations (2.34) can be written as

$$R^I_\mu - \frac{1}{2}(R+\lambda)e^I_\mu = 8\pi G\ T^I_\mu. \tag{2.37}$$

$T^I_\mu(x)$ is called the energy-momentum tensor. It is the sum of the individual energy momentum tensors of the various matter term.[6]

Particles. The trajectory $x^\mu(s)$ of a point particle is an approximate notion. Macroscopic objects have finite size and elementary particles are quantum entities and therefore have no trajectories. At macroscopic scales, the notion of a point particle trajectory is nevertheless very useful.

In the absence of other forces, the equations of motion for the worldline $\gamma : s \mapsto x^\mu(s)$ of a particle are determined by the action

$$S[e,\gamma] = m \int ds\ \sqrt{-\eta_{IJ}\ v^I(s)\ v^J(s)} \tag{2.38}$$

where

$$v^I(s) = e^I_\mu(x(s))\ v^\mu(s) \tag{2.39}$$

and v^μ is the particle velocity

$$v^\mu(s) = \dot{x}^\mu(s) \equiv \frac{dx^\mu(s)}{ds}. \tag{2.40}$$

This action is independent from the way the trajectory is parametrized, and therefore it determines the path, not its parametrization. With the parametrization choice $v_I v^I = -1$, the equations of motion are

$$\ddot{x}^\mu = \Gamma^\mu_{\nu\rho}\ \dot{x}^\nu \dot{x}^\rho, \tag{2.41}$$

where

$$\Gamma^\sigma_{\mu\nu} \quad = \quad e^\rho_J e^{J\sigma}(e_{\rho I}\partial_{(\mu}e^I_{\nu)} + e_{\nu I}\partial_{[\mu}e^I_{\rho]} + e_{\mu I}\partial_{[\nu}e^I_{\rho]}). \tag{2.42}$$

is called the Riemannian or Christoffel connection. In an arbitrary parametrization the equations of motion are

$$\ddot{x}^\mu - \Gamma^\mu_{\nu\rho}\ \dot{x}^\nu \dot{x}^\rho = I(s)\ \dot{x}^\mu \tag{2.43}$$

where $I(s)$ is an arbitrary function of s.

Minkowski solution. Consider a regime in which we can assume that the Newton constant G is small, that is, a regime in which we can neglect the effect of matter on the gravitational field. Assume also that, within our approximation, the cosmological constant λ is negligible. The Einstein equations admit then (among many others) the particularly interesting solution

$$e^I_\mu(x) = \delta^I_\mu, \qquad \omega^I_{\mu J}(x) = 0 \tag{2.44}$$

which is called the Minkowski solution. This solution is everywhere flat.

[6] The energy momentum tensor defined as the variation of the action with respect to the gravitational field may differ by a total derivative from the one conventional in particle physics defined as the Noether current of translations.

Assume that the gravitational field is in this configuration. What are the equations of motion of the matter interacting with this particular gravitational field? These are easily obtained by inserting the Minkowski solution (2.44) into the matter action

$$S[A, \varphi, \psi] = S[e = \delta, \omega = 0, A, \varphi, \psi]. \tag{2.45}$$

The action $S[A, \varphi, \psi]$ is the action of the standard model used in high energy physics. This action is usually written in terms of the spacetime Minkowski metric $\eta_{\mu\nu}$. This metric is obtained from the Minkowski value (2.44) of the tetrad field. For instance, in the action of a scalar field (2.29) the combination $\eta^{IJ}\ e^{\mu}_{I}(x)\ e^{\nu}_{J}(x)$ becomes

$$\eta^{IJ}\ e^{\mu}_{I}(x)\ e^{\nu}_{J}(x) = \eta^{IJ}\ \delta^{\mu}_{I}\ \delta^{\nu}_{J} = \eta^{\mu\nu} \tag{2.46}$$

on this solution.

The Minkowski metric $\eta_{\mu\nu}$ of special relativistic physics is nothing but a particular value of the gravitational field. It is one of the solutions of the Einstein equation, within a certain approximation.

2.1.3 Gauge invariance

The general definition of a system with a gauge invariance, and the one which is most useful for understanding the physics of gauge systems is the following, due to Dirac. Consider a system of evolution equations in an evolution parameter t. The system is said to be "gauge" invariant if evolution is under-determined. That is, if there are two distinct solutions that are equal for t less than a certain $\hat{t}$. See Figure 2.1. These two solutions are said to be "gauge equivalent". Any

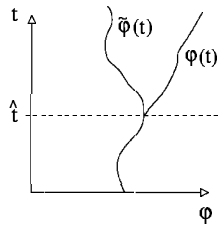


Figure 2.1: Dirac definition of gauge: two different solutions of the equations of motion must be considered gauge equivalent if they are equal for $t < \hat{t}$.

two solutions are said to be gauge equivalent if they are gauge equivalent (as above) to a third solution. The gauge group $\mathcal{G}$ is a group that acts on the physical fields and maps gauge equivalent solutions into one another. Since classical physics is deterministic, under-determined evolution equations are physically consistent only under the stipulation that only quantities invariant under gauge transformations are physical predictions of the theory. These quantities are called the gauge invariant observables.

The equations of motion derived by the action (2.33) are invariant under three groups of gauge transformations: (i) local Yang-Mills gauge transformations, (ii) local Lorentz transformations and (iii) diffeomorphism transformations. They are described below. Gauge invariant observables must be invariant under these three groups of transformations.

(i) Local G transformations. G is the Yang-Mills group. A local G transformation is labeled by a map $\lambda : M \to$

G. It acts on φ, ψ and the connection A in the well known form, while e and ω are invariant

$$\begin{aligned}
\lambda: \quad \varphi(x) &\mapsto R_\varphi(\lambda(x))\ \varphi(x), && (2.47)\\
\psi(x) &\mapsto R_\psi(\lambda(x))\ \psi(x), && (2.48)\\
A_\mu(x) &\mapsto R(\lambda(x))\ A_\mu(x) + R(\lambda(x))\partial_\mu R^{-1}(\lambda(x)), && (2.49)\\
e^I_\mu(x) &\mapsto e^I_\mu(x), && (2.50)\\
\omega^I_{\mu J}(x) &\mapsto \omega^I_{\mu J}(x). && (2.51)
\end{aligned}$$

where R_φ and R_ψ are the representations of G to which φ and ψ belong and R is the adjoint representation.

(ii) Local Lorentz transformations. A local Lorentz transformation is labeled by a map $\lambda : M \to SO(3,1)$. It acts on φ, ψ and the connection ω precisely as a Yang-Mills local transformation with Yang-Mills group $G = SO(3,1)$. Scalars φ belong to the trivial representation; fermions ψ belong to the spinor representations S. The gravitational field e transforms in the fundamental representation. Explicitly, writing an element of $SO(3,1)$ as $\lambda^I{}_J$, we have

$$\begin{aligned}
\lambda: \quad \varphi(x) &\mapsto \varphi(x), && (2.52)\\
\psi(x) &\mapsto S(\lambda(x))\ \psi(x), && (2.53)\\
A_\mu(x) &\mapsto A_\mu(x) && (2.54)\\
e^I_\mu(x) &\mapsto \lambda^I{}_J(x)\ e^J_\mu(x), && (2.55)\\
\omega^I_{\mu J}(x) &\mapsto \lambda^I{}_K(x)\omega^K_{\mu\delta}(x)\lambda^L{}_J(x) + \lambda^I_K(x)\partial_\mu\lambda^K{}_J(x). && (2.56)
\end{aligned}$$

(iii) Diffeomorphisms. Third, and most important of all, is the invariance under diffeomorphisms. A diffeomorphism gauge transformation is labeled by a smooth invertible map $\phi : M \to M$ (that is, by a "diffeomorphism" of M)[7]. It acts *nonlocally* on all the fields, by pulling them back, according to their form character: φ and ψ are zero forms, e, ω and A are one-forms:[8]

$$\begin{aligned}
\phi: \quad \varphi(x) &\mapsto \varphi(\phi(x)), && (2.57)\\
\psi(x) &\mapsto \psi(\phi(x)), && (2.58)\\
A_\mu(x) &\mapsto \frac{\partial\phi^\nu(x)}{\partial x^\mu}\ A_\nu(\phi(x)) && (2.59)\\
e^I_\mu(x) &\mapsto \frac{\partial\phi^\nu(x)}{\partial x^\mu}\ e^I_\nu(\phi(x)), && (2.60)\\
\omega^I_{\mu J}(x) &\mapsto \frac{\partial\phi^\nu(x)}{\partial x^\mu}\ \omega^I_{\nu J}(\phi(x)). && (2.61)
\end{aligned}$$

These three groups of transformations send solutions of the equations of motion into solutions of the equations of motion. They are gauge transformations because we can take these transformations to be the identity before a given coordinate time $\hat{t}$ and different from the identity afterwards. Therefore they are responsible for the under-determination of the evolution equations. Following Dirac's argument given above, physical predictions of the theory must be given by quantities invariant under all three these transformations.

In particular, let a local quantity in spacetime be a quantity dependent on a fixed given point x. Notice that such a quantity cannot be invariant under a diffeomorphism. Therefore no local quantity in spacetime (in this sense) is a gauge invariant observable in GR. The meaning of this fact and the far reaching consequences of diffeomorphism invariance are discussed below in section 2.3.2.

---

[7]There is an unfortunate terminological imprecision. A map $\phi : M \to M$ is called a diffeomorphism. The associated transformations (2.57-2.61) on the fields are also often loosely called a diffeomorphism (also in this book), instead of diffeomorphism gauge transformations. This tends to generate confusion.

[8]Under this definition, internal Lorentz and gauge indices do not transform under a diffeomorphism. Alternatively, one should consider fiber preserving diffeomorphisms of the Lorentz and gauge bundle. This alternative can be viewed as mathematically more clean and physically more attractive, because it makes more explicit the fact that local inertial frames or local gauge choices at different spacetime points cannot be identified (see later). However, the mathematical description of a diffeomorphism becomes more complicated, while the two choices are ultimately physically equivalent, due to the gauge invariance under local Lorentz and gauge transformations.

2.1.4 Physical geometry

At each point x of the spacetime manifold M, the gravitational field $e^I_\mu(x)$ defines a map from the tangent space T_xM to Minkowski space. The map sends a vector v^μ in T_xM into the Minkowski vector $u^I = e^I_\mu(x)v^\mu$. The Minkowski length $|u| = \sqrt{-u \cdot u} = \sqrt{-\eta_{IJ}\, u^I u^J}$ defines a norm $|v|$ of the tangent vector v^μ

$$|v| \equiv |u| = \sqrt{-\eta_{IJ}\ (e^I_\mu(x)v^\mu)\ (e^J_\nu(x)v^\nu)}. \tag{2.62}$$

$|v|$ is called the "physical length" of the tangent vector v. The tangent vector v is called timelike (spacelike or lightlike) if u is timelike (spacelike or lightlike).

This fact allows us to assign a size to any d-dimensional surface in M: At any point x of the surface, the gravitational field maps the tangent space of the surface into a surface in Minkowski space. This surface carries a volume form, which can be pulled back to the tangent space of x and then to the surface itself, and integrated. In particular:

The length L of a curve $\gamma : s \mapsto x^\mu(s)$ is the line integral of the norm of its tangent

$$\mathbf{L}[e,\gamma] = \int |d\gamma| = \int ds\ |u(s)| = \int ds\ \sqrt{-\eta_{IJ}\, u^I(s)\, u^J(s)}, \tag{2.63}$$

where

$$u^I(s) = e^I_\mu(\gamma(s))\ \frac{dx^\mu(s)}{ds}. \tag{2.64}$$

This can be written as the the line integral of the norm of the one-form $e^I(x) = e^I_\mu(x)dx^\mu$ along γ:

$$\mathbf{L}[e,\gamma] = \int_\gamma |e| \tag{2.65}$$

The length is independent from the parametrization and the orientation of γ. A curve is called timelike if its tangent is everywhere timelike. Notice that the action of a particle (2.38) is nothing but the length of its path in spacetime

$$S[e,\gamma] = m\ \mathbf{L}[e,\gamma]. \tag{2.66}$$

The area A of a two-dimensional surface $\mathcal{S} : \sigma = (\sigma^i) \mapsto x^\mu(\sigma^i)$, $i = 1, 2$ immersed in M, is

$$\mathbf{A}[e,\mathcal{S}] = \int |d^2\mathcal{S}| = \int_{\mathcal{S}} d^2\sigma\ \sqrt{\det\,(u_i \cdot u_j)} \tag{2.67}$$

where

$$u^I_i(\sigma) = e^I_\mu(\gamma(\sigma))\ \frac{\partial x^\mu(\sigma)}{\partial \sigma^i} \tag{2.68}$$

and the determinant is over the i, j indices. That is

$$\mathbf{A}[e,\mathcal{S}] = \int d^2\sigma\ \sqrt{(u_1 \cdot u_1)(u_2 \cdot u_2) - (u_1 \cdot u_2)^2} \tag{2.69}$$

A surface is called spacelike if its tangents are all spacelike.

The volume V of a three-dimensional region $\mathcal{R} : \sigma = (\sigma^i) \mapsto x^\mu(\sigma^i)$, $i = 1, 2, 3$ immersed in M, is

$$\mathbf{V}[e, \mathcal{R}] = \int |d^3\mathcal{R}| = \int_{\mathcal{R}} d^3\sigma \ \sqrt{n \cdot n} \tag{2.70}$$

where

$$n_I = \epsilon_{IJKL} \ u_1^J u_2^K u_3^L \tag{2.71}$$

is normal to the surface. A region is called spacelike if n is everywhere timelike.

The quantities **L**, **A** and **V** are particular functions of the gravitational field e. The reason they have these geometrical names is discussed below in section 2.2.3.

2.1.5 Holonomy and metric

In GR, quantities close to observations, such as lengths and areas, are nonlocal, in the sense that they depend on finite but extended regions in spacetime, such as lines and surfaces. Another natural nonlocal quantity, which plays a central role in the quantum theory, is the holonomy U of the gravitational connection (ω, or its self dual part A) along a curve γ.

Definition of the holonomy. Given a connection A in a group G over a manifold M, the holonomy is defined as follows. Let a curve γ be a continuous, piecewise smooth map from the interval $[0, 1]$ into M,

$$\gamma : [0, 1] \longrightarrow M \tag{2.72}$$

$$s \longmapsto x^\mu(s) \ . \tag{2.73}$$

The holonomy, or parallel propagator, $U[A, \gamma]$ of the connection A along the curve γ is the element of G defined by

$$U[A, \gamma](0) \ = \ \mathbb{1} \ , \tag{2.74}$$

$$\frac{d}{ds} U[A, \gamma](s) \ + \ \dot\gamma^\mu(s) \, A_\mu\big(\gamma(s)\big) \ U[A, \gamma](s) \ = \ 0 \ , \tag{2.75}$$

$$U[A, \gamma] \ = \ U[A, \gamma](1) \tag{2.76}$$

where $\dot\gamma^\mu(s) \equiv \frac{dx^\mu(s)}{ds}$ is the tangent to the curve. The formal solution of this equation is

$$U[A, \gamma] = \mathcal{P} \exp \int_0^1 ds \ \dot\gamma^\mu(s) \ A_\mu^i\big(\gamma(s)\big) \, \tau_i \ \equiv \ \mathcal{P} \ \exp \int_\gamma A \ , \tag{2.77}$$

where τ_i is a basis in the Lie algebra of the group G and the path ordered $\mathcal{P}$ is defined by the power series expansion

$$\mathcal{P} \exp \int_0^1 ds \, A\big(\gamma(s)\big) \ \ = \ \ \sum_{n=0}^{\infty} \int_0^1 ds_1 \int_0^{s_1} ds_2 \cdots \int_0^{s_{n-1}} ds_n \, A\big(\gamma(s_n)\big) \cdots A\big(\gamma(s_1)\big) \ . \tag{2.78}$$

The connection A is a rule that defines the meaning of parallel transporting a vector in a representation R of G from a point of M to a nearby point: the vector v at x is defined to be parallel to the vector $v + R(A_a dx^\mu)v$ at $x + dx$. A vector is parallel transported along γ to the vector $R(U(A, \gamma))v$.

A technical remark that we shall need later on: The holonomy of any curve γ

is well defined even if there are finite sets of points where γ is non differentiable and A is ill defined. The reason is that we can break γ in components where everything is differentiable and define the holonomy of γ as the product of the holonomies of the components, which are well defined by continuity.

Physical interpretation of the holonomy. Consider two left handed neutrinos that meet at the spacetime point A, separate and then meet again at the spacetime point B. Assume their spin is parallel at A and evolves under the sole influence of the gravitational field. What is their relative spin at B? A left handed neutrino lives in the self-dual representation of the Lorentz group and therefore its spin is parallel transported by the self dual connection A. Let γ_1 and γ_2 be the worldlines of the two neutrinos from A to B and let $\gamma = \gamma_2^{-1} \circ \gamma_1$ be the loop formed by the two worldlines. If the first neutrino has spin ψ at B, the second has spin $\psi' = U(A,\gamma)\psi$. By having the two neutrinos interact, we can in principle measure a quantity such as $\alpha = 2Re\langle\psi|\psi'\rangle$, which (assuming $|\psi| = 1$) gives the trace of the holonomy $\alpha = \mathrm{tr}\, U[A,\gamma]$.

Metric notation. Einstein wrote GR in terms of the metric field. Here I give the translation to metric variables. Notice, however, that this is necessarily incomplete, since the fermion equations of motion cannot be written in terms of the metric field.

The metric field g is a symmetric tensor field defined by

$$g_{\mu\nu}(x) = e^I_\mu(x)\; e^J_\nu(x)\; \eta_{IJ}. \tag{2.79}$$

At each point x of M, g defines a scalar product in the tangent space $T_x M$

$$(u,v) = g_{\mu\nu}(x) u^\mu v^\nu, \qquad u, v \in T_x M. \tag{2.80}$$

and therefore maps $T_x M$ into $T^*_x M$. In other words, $g_{\mu\nu}$ and its inverse $g^{\mu\nu}$ can be used to raise and lower tangent indices. The fact that $e^\mu_I(x) \equiv \eta_{IJ}\; g^{\mu\nu}\; e^J_\nu(x)$ is the inverse matrix of $e^I_\mu(x)$ is then a result, not a definition.

The linear connection Γ is the field $\Gamma^\rho_{\mu\nu}(x)$ defined by

$$\Gamma^\rho_{\mu\nu} = e^\rho{}_I(\partial_\mu e^I_\nu + \omega^I_{\mu J}\; e^J_\nu). \tag{2.81}$$

It defines a covariant partial derivative D_μ on all fields that have tangent (μ) indices

$$D_\mu v^\nu = \partial_\mu v^\nu + \Gamma^\nu_{\mu\rho} v^\rho; \tag{2.82}$$

together with ω, it defines a covariant partial derivative D_μ on all objects that have Lorentz as well as tangent indices. In particular, notice that equation (2.81) yields immediately

$$D_\mu e^I_\nu = \partial_\mu e^I_\nu + \omega^I_{\mu J}\; e^J_\nu - \Gamma^\rho_{\mu\nu}\; e^I_\rho = 0. \tag{2.83}$$

The antisymmetric part $T^\rho_{\mu\nu} = \Gamma^\rho_{\mu\nu} - \Gamma^\rho_{\nu\mu}$ of the linear connection gives the torsion $T^I = e^I_\rho T^\rho_{\mu\nu} dx^\mu dx^\nu$, defined in (2.5).

The Riemannian or Christoffel connection is the linear connection determined by e and $\omega[e]$. That is, it is defined by

$$\partial_\mu e^I_\nu + \omega[e]^I_{\mu J}\; e^J_\nu - \Gamma^\rho_{\mu\nu}\; e^I_\rho = 0 \tag{2.84}$$

whose solution is (2.42). Notice that the antisymmetric part of this equation is the first Cartan structure equation (2.6), which is sufficient to determine $\omega[e]$ as a function of e.

The Christoffel connection is uniquely determined by g: it is the unique torsion free linear connection that satisfies

$$D_\mu g_{\nu\rho} = 0. \tag{2.85}$$

That is

$$\partial_\mu g_{\nu\rho} - \Gamma^\sigma_{\mu\nu} g_{\sigma\rho} - \Gamma^\sigma_{\mu\rho} g_{\nu\sigma} = 0 \tag{2.86}$$

This equation is solved by (2.42), or

$$\Gamma^\rho_{\mu\nu} = \frac{1}{2} g^{\rho\sigma}(\partial_\mu g_{\sigma\nu} + \partial_\nu g_{\mu\sigma} - \partial_\sigma g_{\mu\nu}). \tag{2.87}$$

Notice that equations (2.84,2.87) alow us to write the explicit solution of the GR equation of motion (2.6)

$$\omega[e]^I_{\mu J} = e^\nu_J(\partial_\mu e^I_\nu - \Gamma^\rho_{\mu\nu} e^I_\rho) \tag{2.88}$$

where Γ is given by (2.87) and g by (2.79). Explicitly, this gives, with a bit of algebra

$$\omega[e]^{IJ}_\mu = 2\, e^{\nu[I}\partial_{[\mu} e_{\nu]}{}^{J]} + e_{\mu K}\, e^{\nu I} e^{\sigma J} \partial_{[\sigma} e_{\nu]}{}^K. \tag{2.89}$$

The Riemann tensor can be defined via

$$R^{\mu}{}_{\nu\rho\sigma}\, e^I_\mu = R^I{}_{J\,\rho\sigma}\, e^J_\nu, \tag{2.90}$$

The Ricci tensor is

$$R_{\mu\nu} \;=\; R^I_\mu\, e_{I\nu} \tag{2.91}$$

(R^I_μ is defined in (2.11)), and the energy momentum tensor (see footnote after (2.37))

$$T_{\mu\nu} \;=\; T^I_\mu\, e_{I\nu}. \tag{2.92}$$

Using this, the Einstein equations (2.37) read

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu}(R+\lambda) = 8\pi G T_{\mu\nu}. \tag{2.93}$$

The Minkowski solution is

$$g_{\mu\nu}(x) = \eta_{\mu\nu}, \tag{2.94}$$

where we see clearly that the spacetime Minkowski metric is nothing but a particular value of the gravitational field. With a straightforward calculation, the action (2.10) reads

$$S[g] = \frac{1}{16\pi\, G}\int (R+\lambda)\, \sqrt{-\det g}\; d^4x. \tag{2.95}$$

The matter action cannot be written in metric variables.

Riemann geometry. The tensor g equips the spacetime manifold M with a metric structure: it defines a distance between any two points, and this distance is a smooth function on M. (More precisely, it defines a pseudo-metric structure, as distance can be imaginary.) Riemann studied the structure defined by (M, g), called today a riemannian manifold, and defined the Riemann curvature tensor, as a generalization of Gauss theory of curved surfaces to an arbitrary number of dimensions. Riemann presented this mathematical theory as a general theory of "geometry" that generalizes Euclidean geometry. Einstein utilized this mathematical theory for describing the physical dynamics of the gravitational field. In retrospect, the reason this was possible is because –as understood by Einstein– the Euclidean structure of the physical space in which we live is determined by the local gravitational field. Therefore elementary physical geometry is simply a description of the local properties of the gravitational field, as revealed by matter (rigid bodies) interacting with it. This point is discussed in more detail below in section 2.2.3.

The basic equations of GR presented in this section do not look too different from the equations of a prerelativistic[9] field theory, such as QED or the standard model. But the similarity can be very misleading. The physical interpretation of a general relativistic theory is very different from the interpretation of a prerelativistic one. In particular, the meaning of the coordinates x^μ is different than in prerelativistic physics, and the gauge invariant observables are not related to the fields as they are in prerelativistic physics.

Understanding the physical meaning of the GR formalism has been a long process. This process has taken decades and perhaps it is not entirely concluded yet. For several decades after Einstein's discovery of the theory, for instance, it was not clear whether or not the theory predicted gravitational waves. The prevailing opinion was that wave solutions were only a coordinate artifact and did not represent physical waves capable of carrying energy and momentum, or, as Bondi put it, capable of "boiling a glass of water". This opinion was wrong of course. Einstein himself got badly wrong on the meaning of the Schwarzshild singularity. Wrong high precision measurements of the Earth-Moon distance have been in the literature for a while, because of a mistake due to a conceptual confusion between physical and coordinate distance...

I do not want to give the impression that GR is foggy. On the very contrary, the fact that in all these and similar instances consensus has eventually emerged indicates that the conceptual

[9] Recall that in this book "relativistic" means *general* relativistic.

structure of GR is secure. But to understand this conceptual structure, to understand how to use the equations of GR correctly and how to relate the quantities appearing in these equations to the numbers measured in the laboratory or observed by the astronomers, is definitely a non-trivial problem. More generally, the problem is to understand what precisely GR says about the world. On the other hand, clarity is this respect is essential if we want to understand the quantum physics of the theory.

In order to shed light on this problem, it is illuminating to retrace the conceptual path and the problems that led to the discovery of the theory. This is done in the following section 2.2. The impatient reader may skip section 2.2 and jump to section 2.3, where the interpretation of GR is compactly presented (but impatience slows understanding...)

2.2 The conceptual path to the theory

The roots of GR are in two distinct problems. Einstein's genius was to understand that the two problems solve each other.

2.2.1 Einstein's 1st problem: A field theory for the Newtonian interaction

It was Newton who discovered dynamics. But to a large extent it was Descartes who, a generation earlier, fixed the general rules of the modern science of nature, or the *Scientia Nova*, as it was called at the time. One of Descartes' prescriptions was the elimination of all the "influences from far away" that plagued mediaeval science. According to Descartes, physical interactions happen only between contiguous entities – as in collisions, pushes and pulls. Newton violated this prescription, describing gravity as the instantaneous "action-at-a-distance" of the force

$$F = G\,\frac{m_1\,m_2}{d^2}. \tag{2.96}$$

Newton did not introduce action-at-a-distance with light heart. He calls it "repugnant". His violation of the Cartesian prescriptions was one of the reasons for the strong initial opposition to Newtonianism. For many, his law of gravitation sounded too much like the discredited "influences from the stars" of the ineffective science of the Middle Ages. But the empirical success of Newton's dynamics and gravitational theory was so immense, that most worries about action-at-a-distance dissipated.

Two centuries later, it is another Briton who finds the way to address the problem afresh, in the effort of understanding electric and magnetic forces. Faraday introduces a new notion[10], which is going to revolutionize modern physics: the notion of *field*. For Faraday, the field was set of lines filling up space. The Faraday lines begin and end on charges; in the absence of charges, each line closes, forming a *loop*. In his wonderful book, which is one of the pillars of modern physics and has virtually no equations, Faraday discusses whether the field is a real physical entity.[11] Maxwell formalizes Faraday's powerful physical intuition into a beautiful mathematical theory –a field theory. At each spacetime point, Maxwell electric and magnetic fields represent the tangent to the Faraday line. There is no action-at-a-distance in the theory: the Coulomb description of the electric force between two charges, namely the instantaneous action-at-a-distance law

$$F = k\,\frac{q_1\,q_2}{d^2}, \tag{2.97}$$

is understood to be correct only in the static limit. A charge q_1 at distance d from another charge q_2 does not produce an instantaneous force on q_2, because if we move q_1 rapidly away, it takes a

[10] Many ideas of modern science have been resuscitated from Hellenistic science [52]. Is the Faraday-Maxwell notion of field a direct descendent of the notion of $\pi\nu\epsilon\upsilon\mu\alpha$, that appears for instance in Hipparchus as the carrier of the attraction of the Moon on the oceans, which causes the tides, as well as in contexts related to magnetism [53]? Did Faraday know this notion?

[11] "With regards to the great point under consideration, it is simply: whether the lines of force have a physical existence or not. ... I think that the physical nature of the lines must be granted."[54] Strictly speaking, we can translate the problem in modern terms as whether the field has degrees of freedom independent from the charges or not. But this doesn't diminishes the ontological valence of Faraday's question, which seems to me transparent in these lines. Faraday's continuation is lovely: "And though I should not have raised the argument unless I had thought it both important and likely to be answered ultimately in the affirmative, I still hold the opinion with some hesitation, with as much, indeed, as accompanies any conclusion I endeavor to draw respecting points in the very depths of science". I think that Faraday's greatness shines in this "hesitation", which betrays his full awareness of the importance of the step he is taking (virtually all of modern fundamental physics comes out of these lines) as well as the full awareness of the risk of taking any major novel step.

time $t = d/c$ before q_2 begins to feel any change. This is the time the interaction takes to move across space at a finite speed, in a manner remarkably consistent with Descartes' prescription.

When Einstein studies physics, Maxwell theory is only three decades old. In his writings, Einstein rhapsodizes on the beauty of Maxwell theory and the profound impression it made upon him. Given the formal similarity of the Newton and Coulomb forces (2.96) and (2.97), it is completely natural to suspect that (2.96) is only correct in the static limit as well. It is natural to suspect that the gravitational force is not instantaneous either: if a neutron star, rushing at great speed from the deep sky, smashed away the Sun, it would take a finite time before any effect be felt on Earth. That is, it is natural to suspect that there is a field theory behind Newton theory as well. Einstein set out to find this field theory. GR is what he found.

Special relativity. In fact, the need for a field theory behind Newton law (2.96) is not just *suggested* by the Coulomb-Maxwell analogy: it is indirectly *required* by Maxwell theory. The reason is that Maxwell theory not only eliminated the apparent action-at-a-distance of Coulomb law (2.97), but it also led to a reorganization of the notions of space and time which, in turn, renders *any* action-at-a-distance inconsistent. This reorganization of the notions of space and time is special relativity, a key step towards GR.

In spite of its huge empirical success, Maxwell theory had an apparent flaw if taken as a fundamental theory[12]: it is not Galilean invariant. Galilean invariance is a consequence of the equivalence of inertial frames – at least it had alway been understood as such. Inertial frame equivalence, or the fact that velocity is a relative notion, is one of the pillars of dynamics. The story goes that in the silent halls of Warsaw's University, an old and grave professor stormed out of his office like a madman, shouting "Eureka! Eureka! The new Archimedes is born!", when he saw Einstein's 1905 paper, offering the solution of this apparent contradiction. The way Einstein solves the problem is an example of theoretical thinking at his best. I think it should be kept in mind as an exemplar, when we consider the apparent contradictions between GR and QM.

Einstein *maintains* his confidence in the Galilean discovery that physics is the same in all moving inertial frames and *also* maintains his confidence that Maxwell equations are correct, in spite of the apparent contradiction. He realizes that there is contradiction only because we implicitly hold a *third* assumption. By dropping this third assumption, the contradiction disappears. The third assumption regards the notion of time. It is the idea that it is always meaningful to say which of two distant events, A and B, happens first. Namely that simultaneity is well defined in a manner independent from the observer. Einstein observes that this is a prejudice we have on the structure of reality. We can drop this prejudice and accept the fact that the temporal ordering of distant events may have no meaning. If we do so, the picture returns to be consistent.

The success of special relativity was rapid, and the theory is today widely empirically supported and universally accepted. Still, I do not think that special relativity has really been fully absorbed yet: the large majority of the cultivated people, as well as a surprising high number of theoretical physicists still believe, deep in their heart, that there is something happening "right now" on Andromeda; that there is a single universal time ticking away the life of the Universe. Do you, my reader?

An immediate consequence of special relativity is that action-at-a-distance is not just "repugnant" as Newton felt: it is a nonsense. There is no (reasonable) sense in which we can say that the force due to the mass m_1 acts on the mass m_2 "instantaneously". If special relativity is correct, (2.96) is not just *likely* to be the static limit of a field theory: it *has* to be the static limit of a field theory. When the neutron stars hits the sun, there is no "now" at which the Earth could feel the

[12] rather than as a phenomenological theory of the disturbances of a mechanical ether whose dynamics is still to be found.

effect. The information that the Sun is not anymore there must travel from Sun to Earth across space, carried by an entity. This entity is the gravitational field.

Maxwell → Einstein. Therefore, shortly after having worked out the key consequences of special relativity, Einstein attacks what is obviously the next problem: searching the field theory that gives (2.96) in the static limit. His aim was to do for (2.96) what Faraday and Maxwell did for (1.7). The result in brief is the following, expressed in modern language.

Maxwell's solution to the problem is to introduce the one-form field $A_\mu(x)$.

The force on the particles is

$$\ddot{x}^\mu = eF^\mu{}_\nu \, \dot{x}^\nu, \qquad (2.98)$$

where F is constructed with the first derivatives of A.
A satisfies the (Maxwell) field equations

$$\partial_\mu F^{\nu\mu} = J^\nu, \qquad (2.99)$$

a system of second order partial differential equations for A, with the charge current J^ν as source.
More generally, the field equations can be obtained as Euler-Lagrange equations of the action

$$S[A, \text{matt}] = \tfrac{1}{4}\int F^* \wedge F + S_{\text{matt}}[A, \text{matt}] \qquad (2.100)$$

where F is the curvature of A.

S_{matt} is obtained from the matter action by replacing derivatives with covariant derivatives.

It follows that the source of the field equations is

$$J^\mu = \frac{\delta}{\delta A_\mu} S_{\text{matt}}[A, \text{matt}]. \qquad (2.101)$$

Einstein's solution is to introduce the field $e^I_\mu(x)$, a one-form with value in Minkowski space.
The force on the particles is (eq (2.41))

$$\ddot{x}^\mu = \Gamma^\mu_{\nu\rho} \, \dot{x}^\nu \dot{x}^\rho, \qquad (2.102)$$

where Γ is constructed with the first derivatives of e (Eq (2.42)).
e satisfies the (Einstein) field equations (eq (2.34), here with $\lambda = 0$)

$$R^I_\mu - \frac{1}{2} e^I_\mu R = 8\pi G \, T^I_\mu, \qquad (2.103)$$

a system of second order partial differential equations for e, with the energy momentum tensor T^I_μ as source.
More in general, the field equations can be obtained as Euler-Lagrange equations of the action ((2.33) in second order form)

$$S[e, \text{matt}] = \tfrac{1}{16\pi G}\int e^I \wedge e^J \wedge R^{KL} \epsilon_{IJKL} + S_{\text{matt}}[e, \text{matt}], \qquad (2.104)$$

where R is the curvature of the connection ω compatible with e.
S_{matt} is obtained from the matter action by replacing derivatives with covariant derivatives and the Minkowski metric with the gravitational metric.
It follows that the source of the field equations is (2.34)

$$T^I_\mu = \frac{\delta}{\delta e^I_\mu} S_{\text{matt}}[e, \text{matt}]. \qquad (2.105)$$

The structural similarity between the theories of Maxwell and Einstein theories is evident. However, this is only half of the story.

2.2.2 Einstein's 2nd problem: Relativity of motion

To understand Einstein's 2nd problem, we have to return again to the origin of modern physics. In the western culture there are two traditional ways of understanding what is "space": as an *entity* or as a *relation*.

"Space is an entity" means that space exists also when there is nothing else than space. It exists by itself, and objects move in it. This is the way Newton describes space, and is called absolute space. It is also the way spacetime (rather than space) is understood in special relativity. Although considered since ancient times (in the democritean tradition), this way of understanding space was not the traditional dominant view in western culture. The dominant view, from Aristotle to Descartes, was to understand space as a relation.

"Space is a relation" means that the world is made by physical objects, or physical entities. These objects have the property that they can be in touch with one another, or not. Space is this "touch", or "contiguity", or "adjacency" relation between objects. Aristotle, for instance, defines the spatial location of an object as the set of the objects that surround it. This is relational space.

Strictly connected to these two manners of understanding space, there are two manners of understanding motion.

"Absolute motion." If space is an entity, motion can be defined as going from one part of space to another part of space. This is how Newton defines motion.

"Relative motion." If space is a relation, motion can only be defined as going from the contiguity of one object to the contiguity of another object. This is how Descartes[13] and Aristotle[14] define motion.

For a physicist, the issue is which of these two ways of thinking about space and motion allows a more effective description of the world.

For Newton, space is absolute and motion is absolute.[15] This is a second violation of Cartesianism. Once more, Newton does not take this step with light heart: he devotes a long initial section of the "Principia" to explain the reasons of his choice. The strongest argument in Newton's favor is entirely a posteriori: his theoretical construction works extraordinary well. Cartesian physics was never as effective. But this is not Newton's argument. Newton recurs to empirical evidence, discussing a famous experiment with a bucket.

---

[13] "We can say that movement is the transference of one part of matter or of one body, from the vicinity of those bodies immediately contiguous to it, and considered at rest, into the vicinity of some others", (Descartes, *Principia Philosophiae*, Sec II-25, pg 51) [55].

[14] Aristotle insists that motion is relative. He illustrates the point with the example of a man walking over a boat. The man moves with respect to the boat, which moves with respect to the water of the river, which moves with respect to the ground... Aristotle's relationalism is tempered by the fact that there are preferred objects that can be used as preferred reference: the Earth at the center of the universe and the celestial spheres, in particular the one of the fixed stars. Thus, we can say that something is moving "in absolute terms" if it moves with respect to the Earth. However, there are *two* preferred frames in ancient cosmology: the Earth and the fixed stars, and the two rotate with respect to each other. The thinkers of the middle ages did not miss this point, and discussed at length whether the stars rotate around the Earth or the Earth rotates under the stars. Remarkably, in the XIVth century Buridan concluded that neither view is more true than the other on ground of reason and Oresme studied the rotation of the Earth, more than a century before Copernicus.

[15] "So, it is necessary that the definition of places, and hence local motion, be referred to some motionless thing such as extension alone or *space*, in so far as space is seen truly distinct from moving bodies", (Newton *De gravitatione et Aequipondio Fluidorum* 89-156) [56]. This is in open contrast with Descartes definition, given in the footnote above.

Newton's bucket. Consider a "bucket full of water, hang by a long cord, so often turned about that the cord is strongly twisted". Whirl the bucket, so that it starts rotating and the cord untwisting. At first

(i) *the bucket rotates (with respect to us) and the water remains still. The surface of the water is flat.*

Then the motion of the bucket is transmitted to the water by friction and thus the water starts rotating together with the bucket. At some time

(ii) *the water and the bucket rotate together. The surface of the water is not anymore flat: it is concave.*

We know from experience that the concavity of the water is caused by rotation. Rotation with respect to what? Newton's bucket experiments shows something subtle about this question. If motion is change of place with respect to the surrounding objects, as Descartes demands, then we must say that in (i) water rotates (with respect to the bucket, which surrounds it), while in (ii) water is still (with respect to the bucket). But, observes Newton, the concavity of the surface appears in (ii), not in (i). It appears when the water is still with respect to the bucket, not when the water moves with respect to the bucket. Therefore the rotation that produces the physical effect is not the rotation with respect to the bucket. It is the rotation with respect to ... what?

It is rotation with respect to space itself, answers Newton. The concavity of the water surface is an effect of the absolute motion of the water: the motion with respect to absolute space, not to the surrounding bodies. This, claims Newton, proves the existence of absolute space.

Newton's argument is subtle, and for three centuries, nobody has been able to defeat it. To understand it correctly, we should lay to rest a common misunderstanding. Relationalism, namely the idea that motion can be defined only in relation to other objects, should not be confused with Galilean relativity. Galilean relativity is the statement that "rectilinear uniform motion" is a priori indistinguishable from stasis. Namely that velocity (just velocity!), is only relative to other bodies. Relationalism, on the other hand, holds that *any* motion (however zigzagging) is a priori indistinguishable from stasis. The very formulation of Galilean relativity assumes a nonrelational definition of motion: "rectilinear and uniform" with respect to what?

Now, when Newton claimed that motion with respect to absolute space is real and physical, he, in a sense, over did it, insisting that even rectilinear uniform motion is absolute. This caused a painful debate, because there are no physical effects of inertial motion, and therefore the bucket argument fails for this particular class of motions.[16] Therefore inertial motion and velocity are to be considered relative in Newtonian mechanics.

What Newton needed for the foundation of dynamics –and what we are discussing here– is not the relativity of inertial motion: it is whether *accelerated* motion, exemplified by the rotation of the water in the bucket, is relative or absolute. The question here is not whether or not there is an absolute space with respect to which velocity can be defined. The question is whether or not there is an absolute space with respect to which *acceleration* can be defined. Newton's answer, supported by the bucket argument was positive. Without this answer, Newton's main law

$$\vec{F} = m\vec{a} \tag{2.106}$$

wouldn't even make sense.

Opposition to Newton's absolute space was even stronger than opposition to his action-at-a-distance. Leibnitz and his school argued fierily against Newton absolute motion and Newton's use of absolute acceleration.[17] Doubts never really disappeared all along the centuries and a feeling kept lingering that something was wrong in Newton's argument. Ernest Mach returned on the issue suggesting that Newton's bucket argument could be wrong because the water does not rotate with respect to absolute space, it rotates with respect to the full matter content of the universe. I will

[16] Newton is well aware of this point, which is clearly stated in the Corollary V of the Principia, but he chooses to ignore it in the introduction of the Principia. I think he did this just to simplify his argument, which was already hard enough for his contemporaries.

[17] Leibnitz had other reasons of complaint with Newton. The two were fighting over the priority for the invention of calculus – scientists' frailties remain the same in all centuries.

comment on this idea and its influence on Einstein in Section 2.4.1. But, as for action-at-a-distance, the immense empirical triumph of Newtonianism could not be overcome.

Or couldn't it? After all, in the early XXth century 43 seconds of arc in Mercury's orbit were observed, which Newton's theory didn't seem to be able to account for...

Generalize relativity. Einstein was impressed by Galilean's relativity. The velocity of a single object has no meaning; only velocity of objects with respect to one another is meaningful. Notice that, in a sense, this is a failure of Newton's program of revealing the "true motions". It is a minor, but significant failure. For Einstein, this was a hint that there is something wrong in the Newtonian (and special relativistic) conceptual scheme.

In spite of its immense empirical success, Newton's idea of an absolute space has something deeply disturbing in it. As Leibniz, Mach, and many others emphasized, space is a sort of extrasensorial entity that acts on objects but cannot be acted upon. Einstein was convinced that the idea of such an absolute space was wrong. There can be no absolute space, no "true motion". Only relative motion, and therefore relative acceleration, must be physically meaningful. Absolute acceleration should not enter physical equations. With special relativity, Einstein had succeeded vindicating Galilean relativity of velocities from the challenge of Maxwell theory. He was then convinced that he could vindicate the entire Aristotelian-Cartesian relativity of motion. In Einstein's terms, "the laws of motion should be the same in all reference frames, not just in the inertial frames". Things move with respect to one another, not with respect to an absolute space; there cannot be any physical effect of absolute motion.

According to many contemporary physicists, this is excessive weight given to "philosophical" thinking, which should not play a role in physics. But Einstein's achievements in physics are far more effective than the ones obtained by these physicists.

2.2.3 The key idea

The question addressed in Newton's bucket's experiment is the following. The rotation of the water has a physical effect –the concavity of the water's surface: with respect to *what* does the water "rotate"? Newton argues that the relevant rotation is not the rotation with respect to the surrounding objects (the bucket), therefore it is rotation with respect to absolute space. Einstein's new answer is simple and fulgurating:

The water rotates with respect to a local physical entity: the gravitational field.

It is the gravitational field, not Newton's inert absolute space, that tells objects if they are accelerating or not, if they are rotating or not. There is no inert background entity such as Newtonian space: there are only dynamical physical entities. Among these are the fields. Among the fields is the gravitational field.

The flatness of concavity of the water surface in Newton's bucket is not determined by the motion of the water with respect to absolute space. It is determined by the physical interaction between the water and the gravitational field.

The two lines of Einstein's thinking about gravity (finding a field theory for the Newtonian interaction – getting rid of absolute acceleration) meet here. Einstein's key idea is that Newton has mistaken the gravitational field for an absolute space.

What leads Einstein to this idea? Why should Newtonian acceleration be defined with respect to the gravitational field? The answer is given by the special properties of the gravitational interaction.[18] These can be revealed by a thought experiment called Einstein's elevator. I present below

[18]Gravity is "special" in the sense that Newtonian absolute space is a configuration of the gravitational field. Once we get rid of the notion of absolute space, the gravitational interaction is not anymore particularly special. It is one

a modern and more realistic version of Einstein's elevator argument.

An "elevator" argument: Newtonian cosmology. Here is a simple physical situation that illustrates that inertia and gravity are the same thing. The model is simple, but completely realistic. It leads directly to the physical intuition underlying GR.

In the context of Newtonian physics, consider a Universe formed by a very large spherical cloud of galaxies. Assume that the galaxies are –and remain– uniformly distributed in space, with a time dependent density $\rho(t)$, and that they attract each other gravitationally. Let C be the center of the cloud. Consider a galaxy A (say, ours) at a distance $r(t)$ from the center C. As is well known, the gravitational force on A due to the galaxies outside a sphere of radius r around C, cancels out, and the gravitational force due to the galaxies inside this sphere is the same as the force due to the same mass concentrated in C. Therefore the gravitational force on A is

$$F = -G\frac{m_A\ \frac{4}{3}\pi r^3(t)\ \rho(t)}{r^2(t)}. \tag{2.107}$$

or

$$\frac{d^2r}{dt^2} = -G\frac{4}{3}\pi\ r(t)\rho(t). \tag{2.108}$$

If the density remains spatially constant, it scales uniformly as r^{-3}. That is $\rho(t) = \rho_0 r^{-3}(t)$, where ρ_0 is a constant equal to the density at $r(t) = 1$. Therefore

$$\frac{d^2r}{dt^2} = -\frac{4}{3}\pi G\rho_0\frac{1}{r^2(t)} = -\frac{c}{r^2(t)} \tag{2.109}$$

where

$$c = \frac{4\pi G\rho_O}{3} \tag{2.110}$$

is a a constant. Equation (2.109) is the Friedmann cosmological equation which governs the expansion of the universe. (It is the same one that one obtains from full GR in the spatially flat case.)

In the Newtonian model we are considering, the galaxy C is in the center of the universe and defines an inertial frame, while the galaxy A is not in the center, and is not inertial. Assume that the cloud is so large that its boundary cannot be observed from C nor from A. If you are in one of these two galaxies, how can you tell in which you are? That is, how can you tell whether you are in the inertial reference frame C or in the accelerated frame A?

The answer is, very remarkably, that you cannot. Since the entire cloud expands or contracts uniformly, the picture of the local sky looks uniformly expanding or contracting precisely in the same manner from all galaxies. But you cannot detect if you are in the inertial galaxy C of in the accelerated galaxy A by local experiments either! Indeed, to detect if we are in an accelerated frame we have to observe inertial forces, such as the ones that make the surface of the water of Newton bucket concave. The A frame accelerates at the acceleration

$$\vec{a} = \frac{c}{r^2(t)}\vec{u} \tag{2.111}$$

where $\vec{u}$ is a unit vector pointing towards C. Therefore there is an inertial force

$$\vec{F}_{\text{inertial}} = -\frac{c}{r^2(t)}\vec{u} \tag{2.112}$$

on all moving masses. This is the force that should allow us to detect that the frame is not inertial. However, all masses feel, besides the local forces $\vec{F}_{local}$, also the cosmological gravitational pull towards C

$$\vec{F}_{\text{cosmological}} = \frac{c}{r^2(t)}\vec{u} \tag{2.113}$$

so that their motion in the accelerated A frame is governed by

$$m\vec{a} = \vec{F}_{\text{local}} + \vec{F}_{\text{inertial}} + \vec{F}_{\text{cosmological}} \tag{2.114}$$

$$= \vec{F}_{\text{local}} \tag{2.115}$$

because (2.112) and (2.113) cancel out exactly. Therefore the local dynamics of all masses in A looks precisely as if it was inertial. The parabola of a falling stone in A, seen from the accelerated A frame, looks as a straight line. There is no way of telling if we are the center, and no way of telling if we are inertial or not.

How to interpret this impossibility of detecting the inertial frame? According to Newtonian physics, the dynamics in C or A is completely different. But this difference is not physically

---

of the fields forming the world. But a very different world than the one of Newton and Maxwell.

observable. In the Newtonian conceptual scheme, A is noninertial, there are gravitational forces and inertial forces, but there is a sort of conspiracy that hides both of them. In fact, the situation is completely general: given an experimental accuracy, in a sufficiently small region, inertial and gravitational forces cancel in a free falling reference system.[19] It is clear that there should be a better way of understanding this physical situation, without resorting to all these unobservable forces.

The better way is to drop the Newtonian preferred *global* frame, and to realize each galaxy has its own *local* inertial reference frame. We can define local inertial frame by the absence of observable inertial effects, as in newtonian physics. Each galaxy has then its local inertial frames. These frames are determined by the gravitational force thereat. That is, it is gravity that determines, at each point, what is inertial. Inertial motion is such with respect to the local gravitational field, not with respect to absolute space.

Gravity determines also the way the frames of different galaxies fall with respect to one another. The gravitational field expresses the relation between the various inertial frames. It is the gravitational field that determines inertial motion. Newton's true motion is not motion with respect to absolute space: it is motion with respect to a frame determined by the gravitational field. It is motion relative to the gravitational field. Equation (2.106) governs the motion of objects with respect to the gravitational field.

The form of the gravitational field. Recall that Einstein's problem was to describe the gravitational field. The discussion above indicates that the gravitational field can be viewed as the field that determines, at each point of spacetime, the preferred frames in which motion is inertial. Let us we write the mathematics that expresses this intuition.

Return to the cloud of galaxies. Since we have dropped the idea of a global inertial reference system, let us coordinatize events in the cloud with *arbitrary* coordinates $x = (x^\mu)$. The precise physical meaning of these coordinates is discussed in detail in the next section. Let x^μ_A be coordinates of a particular event A, say in our galaxy. Since these coordinates are arbitrarily chosen, motion described in the coordinates x^μ is in general not inertial in our galaxy. For instance, particles free from local forces do not follow straight lines. But we can find a local inertial reference frame around A. Let us denote the coordinates it defines as X^I, and take the event A as the origin, so that $X^I(A) = 0$. The coordinates X^I can be expressed as functions

$$X^I = X^I(x) \tag{2.116}$$

of the arbitrary coordinates x. In the x coordinates, the non-inertiality of the motion in A is gravity. Gravity in A is the information of the change of coordinates that takes us to inertial coordinates. This information is contained in the functions (2.116). But only the value of these functions in a small neighborhood around A is relevant, because if we move away, the local inertial frame will change. Therefore we can Taylor expand (2.116) and keep only the first nonvanishing term. As $X^I(A) = 0$, to first nonvanishing order we have

$$X^I(x) = e^I_\mu(x_A)\ x^\mu \tag{2.117}$$

[19] This is the equivalence principle. By the way, Newton, the genius, knew it: "If bodies, moved among themselves, are urged in the direction of parallel lines by equal accelerative forces, they will all continue to move among themselves, after the same manner as if they had not been urged by those forces" (Newton, Principia, Corollary VI to the Laws of Motion) [57]. Newton uses this corollary for computing the complicated motion of the Moon in the solar system. In the frame of the Earth, inertial forces and the Solar gravity cancel out with good approximation, and the Moon follows a Keplerian orbit.

where we have defined

$$e^I_\mu(x_A) = \left.\frac{\partial X^I(x)}{\partial x^\mu}\right|_{x=x(A)}. \tag{2.118}$$

The quantity $e^I_\mu(x_A)$ gives all the information we need to know the local inertial frame in A. The construction can be repeated at each point x. The quantity

$$e^I_\mu(x) = \left.\frac{\partial X^I(x)}{\partial x^\mu}\right|_{x} \tag{2.119}$$

where X^I are now inertial coordinates at x, is the gravitational field at x. This is the form of the field introduced in section 2.1.1.

The gravitational field $e^I_\mu(x)$ is therefore the Jacobian matrix of the change of coordinates from the x coordinates to the coordinates X^I that are locally inertial at x. The field $e^I_\mu(x)$ is also called the "tetrad" field, from the Greek word for "four", or the "soldering form", because it "solders" a Minkowski vector bundle to the tangent bundle, or, following Cartan, the "moving frame", although there is nothing moving about it.

Transformation properties. If the coordinate system X^I defines a local inertial system at a given point, so does any other local coordinate system $Y^J = \Lambda^J{}_I X^I$, where Λ is a Lorentz transformation. Therefore the index I of $e^I_\mu(x)$ transforms as a Lorentz index under a local Lorentz transformation, and the two fields $e^I_\mu(x)$ and

$$e'^J{}_\mu(x) = \Lambda^J{}_I(x) e^I_\mu(x) \tag{2.120}$$

represent the same physical gravitational field. Therefore this description of gravity has a local Lorentz gauge invariance.

What happens if instead of using the physical coordinates x we chose coordinates $y = y(x)$? The chain rule determines the field $e'^I_\nu(y)$ that we would have found had we used coordinates y

$$e'^I_\nu(y) = \frac{\partial x^\mu(y)}{\partial y^\nu} e^I_\mu(x(y)). \tag{2.121}$$

The transformation properties (2.120) and (2.121) are precisely the transformation properties (2.55) and (2.60) under which the GR action is invariant.

They are also transformation laws of a one-form field valued in a vector bundle P over the spacetime manifold M, whose fiber is Minkowski space $\mathcal{M}$, associated to a principal $SO(3,1)$ Lorentz bundle. This is a natural geometric setting for the gravitational field. The connection ω defined in section 2.1.1 is a connection of this bundle. This setting realizes the physical picture of a patchwork of Minkowski spaces, suggested by the cloud of galaxies, carrying Lorentz frames at each galaxy. More precisely, the gravitational field can be viewed as map $e : TM \to P$ that sends tangent vectors to Lorentz vectors.

Matter. Finally, consider a particle moving in spacetime along a worldline $x^\mu(\tau)$. If a particle has velocity $v^\mu = dx^\mu/d\tau$ at a point x, its velocity in local Minkowski coordinates X^I at x is

$$u^I = \left.\frac{\partial X^I(x)}{\partial x^\mu}\right|_{x} v^\mu = e^I_\mu(x) v^\mu \tag{2.122}$$

In this local Minkowski frame, the infinitesimal action along the trajectory is

$$dS = m\sqrt{-\eta_{IJ} u^I u^J}\, d\tau. \tag{2.123}$$

Therefore the action along the trajectory is the one given in (2.38). The same argument applies to all matter fields: the action is a sum over spacetime of local terms which can be inferred from their Minkowski space equivalent.

Metric geometry. In section 2.1.4, we have seen that the gravitational field e defines a metric structure over spacetime. One is often tempted to give excessive significance to this structure, as if distance was an essential property of reality. But there is no a priori Kantian notion of distance needed to understand the world. We could have developed physics without ever thinking about distances, and nevertheless retaining the complete predictive and descriptive power of our theories.

What is the physical meaning of the spacetime metric structure? What do we mean when we say that two points are three centimeters apart, or two events are 3 seconds apart?

The answer is in the dynamics of matter interacting with the gravitational field. Let us first consider Minkowski space. Consider two objects A and B that are three centimeters apart. This means that if we put a ruler between the two points, the part of the ruler that fits between the two is marked 3 cm. The shape of the ruler is determined by the Maxwell and Schrödinger equations at the atomic level. These equations contain the Minkowski tensor η_{IJ}. They have stable solutions in which the molecules keep themselves (better: vibrate around equilibrium positions) at fixed "distance" L from one another. L is determined by the constants in these equation. This means that the molecules keep themselves at points with coordinate distances Δx^I's such that

$$\eta_{IJ}\Delta x^I \Delta x^J = L^2. \tag{2.124}$$

We exploit this peculiar behavior of condensed matter for coordinatizing spacetime locations. That is, "distance" is nothing but a convenient manner for labeling locations determined by material objects (the ruler), whose dynamics is governed by certain equations. We could avoid mentioning distance by saying a number $N = 3cm/L$ of molecules, obeying the Maxwell and Shrödinger equations with given initial values, fit between A and B.

Consider now the same situation in a gravitational field e. Again, the fact that two points A and B are three centimeters apart means that we can fit the N molecules of the ruler between A and B. But now the dynamics of the molecules is determined by their interaction with the gravitational field. The Maxwell and Shrödinger equations have stable solutions in which the molecules keep themselves at coordinate distances Δx^μ's such that

$$\eta_{IJ}e^I_\mu(x)e^J_\nu(x)\Delta x^\mu \Delta x^\nu = L^2. \tag{2.125}$$

Thus, a measure of distance is a measurement of the local gravitational field, performed exploiting the peculiar way matter interacts with gravity.

The same is true for temporal intervals. Consider two events A and B that happen one after the other. The meaning that three second have lapsed between A and B is that a clock has ticked three times in this time interval. The physical system that we use as a clock interacts with the gravitational field. The pace of the clock is determined by the local value of e. Thus, a clock is nothing but a device measuring an extensive function of the gravitational field along a worldline going from A to B.

Imagine that a particle falls along a timelike geodesic from A to B. We know from special relativity that the increase of the action of the particle in the particle frame is

$$dS = mdt. \tag{2.126}$$

where m is the particle mass. Therefore a clock commoving with the particle will measure the quantity

$$T = \frac{1}{m}S = \int_A^B d\tau \sqrt{-\eta_{IJ}e^I_\mu e^J_\nu \dot{x}^\mu \dot{x}^\nu} \tag{2.127}$$

Therefore a clock is a device for measuring a function T of the gravitational field. In general any metric measurement is nothing but a measurement of a nonlocal function of the gravitational field.

This is true in an arbitrary gravitational field e as well as in flat space. In flat space, we can use these measurements for determining positions with respect to the gravitational field. Since the flat space gravitational field is Newton absolute space, these measurements locate points in spacetime.

2.2.4 Active and passive diffeomorphisms

Before getting to the last and main step in Einstein's discovery of GR, we need the notion of active diffeomorphism. I introduce this notion with an example.

Consider the surface of the Earth, and call it M. At each point $P \in M$ on Earth, say the city of Paris, there is a certain temperature, say $T(P)$. The temperature is a scalar function $T : M \to R$ on the Earth surface. Imagine a simplified model of weather evolution in which the only factor determining temperature change was the displacement of air due to wind. By this I mean the following. Fix a time interval: say we call T the temperature on May first, and

$\tilde{T}$ the temperature on May 2nd. During this time interval, the winds move the air which is over a point $Q = \phi(P)$ to the point P. If, say, Q is the French village of Quintin, this means that the winds have blown the air of Quintin to Paris. Let then the temperature $\tilde{T}(P)$ of Paris on May 2nd is equal to the temperature $T(Q)$ of Quintin the day before. The "wind" map ϕ is a map from the Earth surface to itself, which associates to each point P the point Q from which the air has been blown by the wind. From May first to May 2nd, the temperature field changes then as follows:

$$T(P) \to \tilde{T}(P) = T(\phi(P)). \tag{2.128}$$

Assuming it is smooth and invertible, the map $\phi : M \to M$ is an *active diffeomorphism*. The scalar field T on M is transformed by this active diffeomorphism as in equation (2.128): it is "dragged" along the surface of the Earth by the diffeomorphism ϕ. Notice that coordinates play no role in all this.

Now, imagine that we choose certain geographical coordinates x to coordinatize the surface of the Earth. For instance, latitude and longitude, namely the polar coordinates $x = (\theta, \varphi)$, with $\varphi = 0$ being Greenwich. Using these coordinates, the temperature is represented by a function of the coordinates $T(x)$. The May first temperature $T(x)$ and the May 2nd temperature $\tilde{T}(x)$ are related by

$$\tilde{T}(x) = T(\phi(x)). \tag{2.129}$$

For instance, if the wind has blown uniformly westward by $2^o20'$ degrees (Quintin is $2^o20'$ west of Paris), then

$$\tilde{T}(\theta, \varphi) = T(\theta, \varphi + 2^o20'). \tag{2.130}$$

Of course, there is nothing sacred about this choice of coordinates. For instance, the French might resent that the origin of the coordinates is Greenwich, and have it pass by Paris, instead. Thus, the French would describe the same temperature field that the British describe as $T(\theta, \varphi)$ by means of different polar coordinates, defined by $\varphi = 0$ being Paris. Since Paris is $2^o20'$ degrees East of Greenwich, for the French, the temperature field on May first is

$$T'(\theta, \varphi) = T(\theta, \varphi + 2^o20'). \tag{2.131}$$

This is a change of coordinates, or a *passive diffeomorphism*.

Now the two equations (2.130) and (2.131) look precisely the same. But it would be silly to confuse them. In equation (2.130), $\tilde{T}(\theta, \varphi)$ is the temperature of May 2nd; while in equation (2.131), $T'(\theta, \varphi)$ is still the temperature of May first, but written in French coordinates. The first equation represents a change in the temperature field due to the wind, the second equation represents a change in convention. The first equation describes an "active diffeomorphism", the second a change of coordinates, also called a "passive diffeomorphism".

Given a manifold M, an active diffeomorphism ϕ is a smooth invertible map from M to M. A scalar field T on M is a map $T : M \to R$. Given an active diffeomorphism ϕ, we define the new scalar field $\tilde{T}$ transformed by ϕ as

$$\tilde{T}(P) = T(\phi(P)). \tag{2.132}$$

Coordinates play no role in this.

A coordinate system x on a d-dimensional manifold M is an invertible differentiable map from (an open set of) M to R^d. Given a field T on M, this map determines the function $t : R^d \to R$ defined by $t(x) = T(P(x))$, called "the field T in coordinates x".[20] A passive diffeomorphism is an invertible differentiable map $\phi : R^d \to R^d$ that defines a new coordinate system x' on M by $x(P) = \phi(x'(P))$. The value of the field T in coordinates x' is given by

$$t'(x') = t(\phi(x')). \tag{2.133}$$

Beware the formal similarity between (2.132) and (2.133).

The above extends immediately to all structures on M. For instance, an active diffeomorphism ϕ carries a one-form field e on M to the new one-form field $\tilde{e} = \phi^* e$, the pull-back of e under ϕ and so on.

In particular, a metric $d : M \times M \to R^+$ is an assignment of a distance $d(A, B)$ between any two points A and B of M. An active diffeomorphism defines the new metric $\tilde{d}$ defined by

[20] In the physics literature, the two maps $T : M \to R$ and $t = x \circ T : R^d \to R$, are always indicated with the same symbol, generating confusion between active and passive diffeomorphisms. In this paragraph I use distinct notations. In the rest of the text, however, I shall adhere to the standard notation and indicate the field and its coordinate representation with the same symbol.

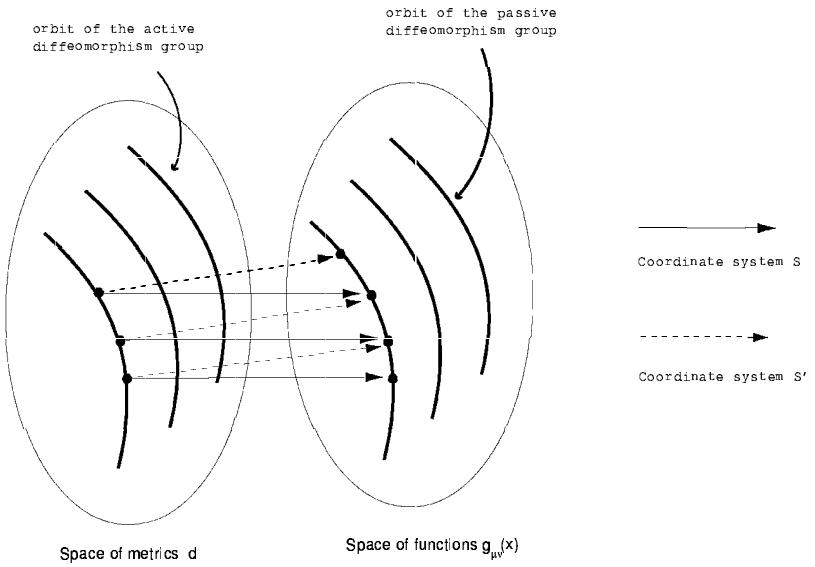


Figure 2.2: Active and passive diffeomorphisms.

$\tilde{d}(A,B) \equiv d(\phi^{-1}(A), \phi^{-1}(B))$. The two metrics d and $\tilde{d}$ are isometric but distinct.[21] An equivalence class of metrics under active diffeomorphisms is sometimes called a "geometry". Given a coordinate system, we can represent a (Riemannian) metric d by means of a tensor field on R^d: Riemann's metric tensor $g_{\mu\nu}(x)$ or, equivalently, the tetrad field $e^I_\mu(x)$. Under a change of coordinate system, the same metric is represented by a different $g_{\mu\nu}(x)$ or $e^I_\mu(x)$.

The example of the Earth temperature given above illustrates a peculiar relation between active and passive diffeomorphisms: given two temperature fields T and $\tilde{T}$ related by an active diffeomorphism, we can always find a coordinate transformation such that in the new coordinates $\tilde{T}$ is represented by the same function as T in the old coordinates. This simple mathematical observation is at the root of Einstein's arguments that I will describe below. (The argument will be essentially that a theory that does not distinguish coordinate systems cannot distinguish fields related by active diffeomorphisms either.)

More precisely, the relation between active and passive diffeomorphisms is as follows. The group of the active diffeomorphisms acts on the space of metrics d. The group of passive diffeomorphisms acts on the space of functions $g_{\mu\nu}(x)$. The orbits of first group are in natural one-to-one correspondence with the orbits of the second. However, the relation between the individual metrics d and the individual functions $g_{\mu\nu}(x)$ depends on the coordinate system chosen. The situation is illustrated in Figure 2.2.

[21] Here is an example of isometric but distinct metrics: The 2001 Shell road map says that the distances between New-York (NY), Chicago (C) and Kansas City (KC) are: d(NY,C)=100miles, d(C,KC)=50miles, d(KC,NY)=100miles, while the 2002 Lonely Planet tourist guide claims that these distances are $\tilde{d}$(NY,C)=100miles, $\tilde{d}$(C,KC)=100miles, $\tilde{d}$(KC,NY)=50miles. Obviously these are not the same distances. But they are isometric: the two are transformed into each other by the active diffeomorphism ϕ(NY)=C, ϕ(C)=KC, ϕ(KC)=NY.

2.2.5 General covariance

Around 1912, using the idea that any motion is relative, Einstein had found the form of the gravitational field as well as the equations of motions of matter in a given gravitational field. This was already a remarkable achievement, but the field equations for the gravitational field were still missing. In fact, the best part of the story had yet to come.

Two problems remained open: the field equations and understanding the physical meaning of the coordinates x^μ introduced above. Einstein struggled with these two problems during the years 1912-1915, trying several solutions and changing his mind repeatedly. Einstein has called this search his "struggle with the meaning of the coordinates". The struggle was epic. The result turned out to be amazing. In Einstein words, it was "beyond my wildest expectations".

To increase Einstein's stress, Hilbert, probably the greatest mathematician at the time, was working on the same problem, trying to be first to find the gravitational field equations. The fact that Hilbert, with his far superior mathematical skills, could not find these equations first, testifies how much fundamental physical problems are profoundly different than mathematical problems.

In his search for the field equations, Einstein was guided by several pieces of information. First, the static limit of the field equations must yield Newton law, as the static limit of Maxwell theory yields Coulomb law. Second, the source of Coulomb law is charge; and the charge density is the temporal component of the four-current $J^\mu(x)$, which is the source of Maxwell equations. The source of the Newtonian interaction is mass. Einstein had understood with special relativity that mass is in fact a form of energy and that the energy density is the temporal component of the the energy-momentum tensor $T_{\mu\nu}(x)$. Therefore $T_{\mu\nu}(x)$ had to be the likely source of the field equations. Third, the introduction of the gravitational field was based on the use of arbitrary coordinates, therefore there should be some of form of covariance under arbitrary changes of coordinates in the field equations. Einstein searched for covariant second order equations as relations between tensorial quantities, since these are unaffected by coordinate change. He learned from Riemannian geometry that the only combination of second derivatives of the gravitational field that transforms tensorially is the Riemann tensor $R^\mu{}_{\nu\rho\sigma}(x)$. This was in fact Riemann's major result. Einstein knew all this in 1912. To derive Einstein's field equations (2.93) from these ideas is a simple calculation, presented in all GR textbooks, which a good graduate student can today repeat easily. Still, Hilbert couldn't do it, and Einstein got stuck for several years. What was the problem?

The problem was "the meaning of the coordinates". Here is the story.

1. Einstein for general covariance. At first, Einstein, demands the field equations for the gravitational field $e^I_\mu(x)$ to be generally covariant on M. This means that if $e^I_\mu(x)$ is a solution, then $e'^I_\nu(y)$ defined in equation (2.121) should also be a solution. For Einstein, this requirement (unheard at the time), was the formalization of the idea that the laws of nature must be the same in all reference frames, and therefore in all coordinate systems.

2. Einstein against general covariance. In 1914, however, Einstein convinces himself that the field equations should *not* be generally covariant [58]. Why? Because Einstein rapidly understands the physical consequences of general covariance, and at first he panics in front of them. The story is very instructive, because it reveals the true magics hidden inside GR. Einstein's argument *against* general covariance is the following.[22]

Consider a region of spacetime containing two spacetime point A and B. Let e be a gravitational field in this region. Say that around the point A the field is flat, while at the point B it is not (see

[22] At first, Einstein got discouraged about generally covariant field equation because of a mistake he was making in deriving the static limit: the calculation yielded the wrong limit. But this is of little importance here, given the powerful use that Einstein has been routinely capable of making of general conceptual arguments.

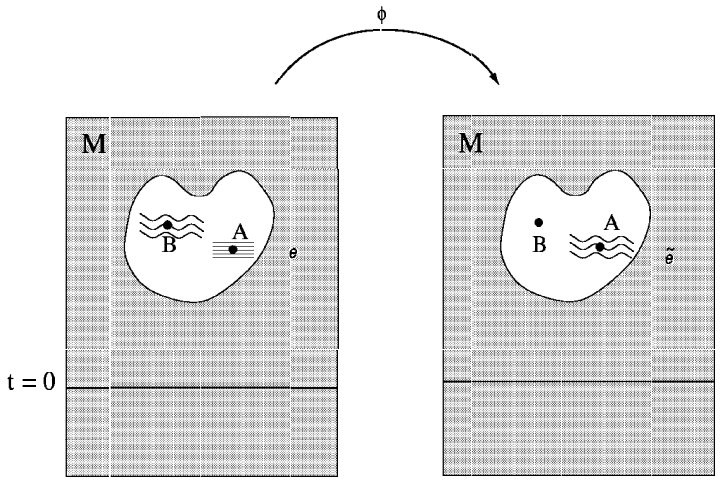


Figure 2.3: The active diffeomorphism ϕ drags the non flat (wavy) gravitational field from the point B to the point A.

Figure 2.3). Next, consider a map ϕ from M to M that maps the point A to the point B. Consider the new field $\tilde{e} = \phi^* e$ which is pulled back by this map. The value of the field $\tilde{e}$ at A is determined by the value of e at B, and therefore the field $\tilde{e}$ will not be flat around A (see Figure 2.3).

Now, if e is a solution of the equations of motion, and if the equations of motion are generally covariant, then $\tilde{e}$ is also a solution of the equations of motion. This is because of the relation between active diffeomorphisms and changes of coordinates: we can always find two different coordinate systems on M, say x and y, such that the function $e^I_\mu(x)$ that represents e in the coordinate system x is the same function as the function $\tilde{e}^I_\mu(y)$ that represents $\tilde{e}$ in the coordinate systems y. Since the equations of motions are the same in the two coordinate systems, the fact that this function satisfies the Einstein equations implies at the same time that e as well as $\tilde{e}$ are physical solutions.

Let me repeat the argument in a different form. We have found in the previous section that if $e^I_\mu(x)$ is a solution of the Einstein equations, then so is $e'^I_\nu(y)$ defined in equation (2.121). But the function e'^I_ν can be interpreted in two distinct manners. First, as the same field as e, expressed in a different coordinate system. Second, as a *different* field, expressed in the same coordinate system. That is, we can *define* the new field as

$$\tilde{e}^I_\mu(x) = e'^I_\mu(x). \tag{2.134}$$

This new field $\tilde{e}$ is genuinely different from e. In general, it will not be flat around A. In particular, the scalar curvature $\tilde{R}$ of $\tilde{e}$ at A is

$$\tilde{R}|_A = \tilde{R}(x_A) = R(\phi(x_A)) = R|_B. \tag{2.135}$$

In other words, if the equations of motion are generally covariant they are *also* invariant under active diffeomorphisms.

Given this, Einstein makes the following famous observation.

The "hole" argument: Assume the gravitational field equations are generally covariant. Consider a solution of these equations in which the gravitational field is e and there is a region H of the universe without matter (the "hole", represented as the white region in Figure 2.3).

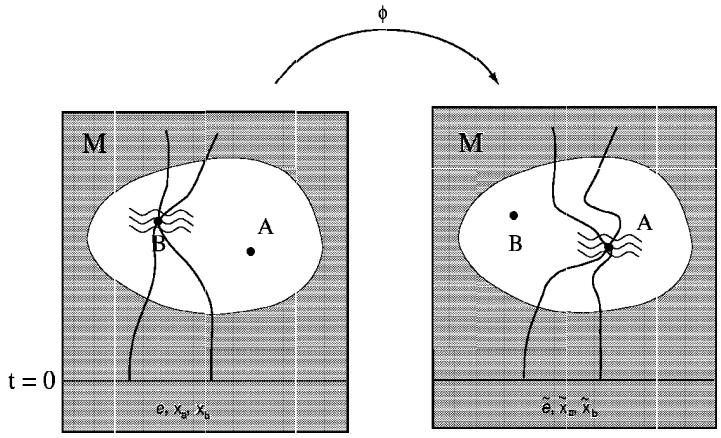


Figure 2.4: The diffeomorphism moves the nonflat region *as well as the intersection point of the two particles* a and b from the point B to the point A.

Assume that inside H there is a point A where e is flat and a point B where it is not flat. Consider a smooth map $\phi : M \to M$ which reduces to the identity outside H, and such that $\phi(A) = B$, and let $\tilde{e} = \phi^* e$ be the pull back of e under ϕ. The two fields e and $\tilde{e}$ have the same past, are both solutions of the field equations, but have different properties at the point A. Therefore the field equations do not determine the physics at the spacetime point A. Therefore they are not deterministic. But we know that (classical) gravitational physics is deterministic. Therefore either

(i) the field equations must not be generally covariant, or

(ii) there is no meaning in talking about the physical spacetime point A.

On the basis of this argument, Einstein searched for non generally covariant field equations for three years, in a frantic race against Hilbert.

3. Einstein return to general covariance. Then, rather suddenly, in 1915, Einstein published generally covariant field equations. What had happened? Why Einstein changed his mind? Is there a mistake in the hole argument? No, the hole argument is correct. The correct physical conclusion, however, is *(ii)*, not *(i)*. Einstein saw this point as a fulguration, as the precise conceptual discovery to which all his previous thought converged.

Einstein's way out from the difficulty raised by the hole argument is to realize that there is no meaning in referring to "the point A" or "the event A", without further specifications.

Let us follow Einstein's explanation in detail.

Spacetime coincidences. Consider again the solution e of the field equations, but assume that in the universe there are also the two particles a and b. Say that the worldlines $(x_a(\tau_a), x_b(\tau_b))$ of the two particles intersect at the spacetime point B (See Figure 2.4).

Now, for given initial conditions, the worldlines of the particles are determined by the gravitational field. They are geodesics of e or, if other forces are involved, they satisfy the geodesic equation with additional force term. Consider the field $\tilde{e} = \phi^* e$. The particles' worldlines $(x_a(\tau_a), x_b(\tau_b))$ are *not* anymore solutions of the particles' equation of motion in this gravitational field. If the gravitational field is $\tilde{e}$ instead of e, the particles' motion over M will be different. But it is easy to

find the motion of the particles determined by $\tilde{e}$, precisely because the complete set of equations of motion is generally covariant. Therefore, an active diffeomorphism on the gravitational field *and* the particles sends solutions into solutions. Therefore, the motion of the particles in the field $\tilde{e}$ is given by the worldlines

$$\tilde{x}_a(\tau_a) = \phi^{-1}(x_a(\tau_a)), \qquad \tilde{x}_b(\tau_b) = \phi^{-1}(x_b(\tau_b)), \tag{2.136}$$

Then the particles a and b do not intersect anymore in B. They intersect in $A = \phi^{-1}(B)$!

Now, instead of asking whether or not the field is flat in A, let as ask whether or not the field is flat in the point where the particles meet.Clearly the result is the same for the two cases (e, x_a, x_b) and $(\tilde{e}, \tilde{x}_a, \tilde{x}_b)$. Formally, assuming the intersection point is at $\tau_a = \tau_b = 0$,

$$\begin{aligned} \tilde{R}|_{inters} &= \tilde{R}(\tilde{x}_a(0)) = R(\phi(\tilde{x}_a(0))) \\ &= R(\phi(\phi^{-1}(x_a(0)))) = R(x_a(0)) = R|_{inters}. \end{aligned} \tag{2.137}$$

This prediction is deterministic. There are no two contradictory predictions, therefore there is determinism, as far as we restrict to this kind of predictions. Einstein calls "spacetime coincidences" this way of determining points.

Einstein observes that this conclusion is general: the theory does not predict what happens at spacetime points (like Newtonian and special relativistic theories do). Rather, it predicts what happens at locations *determined by the dynamical elements of the theory themselves.* In Einstein's words:

> "All our space-time verifications invariably amount to a determination of space-time coincidences. If, for example, events consisted merely in the motion of material points, then ultimately nothing would be observable but the meeting of two or more of these points. Moreover, the results of our measuring are nothing but verifications of such meetings of the material points of our measuring instruments with other material points, coincidences between the hands of a clock and points on the clock dial, and observed point-events happening at the same place at the same time. The introduction of a system of reference serves no ther purpose than to facilitate the description of the totality of such coincidences."[59]

The two solutions (e, x_a, x_b) and $(\tilde{e}, \tilde{x}_a, \tilde{x}_b)$ are only distinguished by their localization on the manifold. They are different in the sense that they ascribe different properties to manifold points. However, if we demand that localization is defined only with respect to the fields and particles themselves, then there is nothing that distinguishes the two solutions physically. In fact, concludes Einstein, the two solutions represent the same physical situation. The theory is gauge invariant, in the sense of Dirac, under active diffeomorphisms: there is a redundancy in the mathematical formalism: the same physical world can be described by different solutions of the equations of motion.

It follows that localization on the manifold has no physical meaning. The physical picture is completely different from the example of the temperature field on the Earth surface illustrated in the previous section. In that example, the cities of Paris and Quintin were real distinguishable entities, independently from the temperature field. In GR, instead, general covariance is compatible with determinism only assuming that individual spacetime points have no physical meaning by themselves. It is like having only the temperature field, without the underlying Earth.

What disappears in this step is precisely the background spacetime that Newton believed to have been able to detect, with great effort, beyond the apparent relative motions.

Einstein's step towards a profoundly novel understanding of nature is achieved. Background space and spacetime are effaced from this new understanding of the world. Motion is entirely

relative. Active diffeomorphism's invariance is the key to implement this complete relativisation. Reality is not made by particles and fields on a spacetime: it is made by particles and fields (including the gravitational field), that can only be localized with respect to one another. No more fields on spacetime: just fields on fields. Relativity has become general.

2.3 Interpretation

General covariance makes the relation between formalism and experiment far more indirect than in conventional field theories.

Take Maxwell theory as an example. We assume that there is a background spacetime. We have special objects at our disposal (the walls of the lab, the Earth) that define an inertial frame to a desired approximation. These objects allow us to designate locations relative to background spacetime. We have two kinds of measuring devices: *(a)* meters and clocks that measure distance and time intervals from these reference objects, and *(b)* devices that measure the electric and magnetic fields. The reading of the devices *(a)* gives us x^μ. The reading of the devices *(b)* gives us $F_{\mu\nu}$. We measure the two and say that the field has the value $F_{\mu\nu}$ at the point x^μ. The theory can predict the value $F_{\mu\nu}$ at the point x^μ.

We cannot do the same in GR. The theory does not predict the value of the field at the point x^μ. So, how do we compare theory and observations?

2.3.1 Observables, predictions and coordinates

As discussed at the end of the previous section, a physical state does not correspond to a solution $e(x)$ of the Einstein's equations, but to an equivalence class of solutions under active diffeomorphisms. Therefore the quantities that the theory predicts are all and only the quantities that are well defined on these equivalence classes. That is, only the quantities that are invariant under diffeomorphisms. These quantities are independent from the coordinates x^μ.

In concrete applications of the theory, these quantities are generally obtained by solving away the coordinates x from solutions of the equations of motion. Here are a few examples.

Solar system. Consider the dynamics of the solar system. The variables are the gravitational field $e(x)$ and the worldlines of the planets, $x_n(\tau_n)$. Fix a solutions $(e(x), x_n(\tau_n))$ of the equations of motion. We want to derive physical predictions from this solution and compare them with observations. Choose for simplicity $\tau_n = x^0$, so that the solution is expressed by $(e(x), \vec{x}_n(x^0))$. Consider the worldline of the Earth. Compute the distance $d_n(x^0)$ between the Earth and the planet n, defined as the proper time lapsed along the Earth worldline while a null geodesic (a light pulse) leaving the Earth at x_0, travels from Earth to the planet and back. The functions $(d_n(x^0))$ can be computed from the given solutions of the equation of motion. Consider a space $\mathcal{C}$ with coordinates (d_n). The functions $(d_n(x^0))$ define a curve γ on this space.

We can associate a measuring device to each d_n: a laser apparatus that measures the distance from other planets. These quantities can be measured together. We obtain the event (d_n), which can be represented by a point in $\mathcal{C}$. The theory predicts that this point will fall on the curve γ. A sequence of these events can be compared with the curve γ, and in this way we can test the given solutions of the equations of motion against experience. (In the terminology of chapter 3, the quantities d_n are partial observables.) Notice that this can be done with arbitrary precision and that distant stars, inertial systems, preferred coordinates, or choice of time variable, play no role.

Clocks. Consider the gravitational field around the Earth. Consider two worldlines. Let the first be the worldine of an object fixed on the Earth surface. Let the second be the worldline of an object in free fall on a Keplerian orbit around the Earth, that is, a satellite. Fix an arbitrary initial point P on the worldline of the orbiting object and let T_1 be the proper time from P along this worldline. Send a light signal from P to the object on Earth, let Q be the point of this worldline when the signal is received and let T_2 be the proper time from Q along this worldline. Let then $T_2(T_1)$ be the reception proper time on Earth of a signal sent at T_1 proper time in orbit. GR allows us to compute the function $T_2(T_1)$ for any T_1.

It is easy to associate measuring devices to T_1 and T_2: these are a clock on Earth and a clock in orbit. If the orbiting object sends a signal at fixed proper times T_1, the reception times T_2 can be compared with the predictions of the theory. Here T_1 and T_2 are the partial observables. I let you decide which one of the two is the "true time variable".

Solar system with a clock. We can add a clock to the Solar system measurements described above. Fixing arbitrarily an initial event on Earth (a particular eclipse, the birth of Jesus or the death of John Lennon), and compute the proper time $T(x^0)$ lapsed from this event along the Earth worldline. The partial observable T can be added to the partial observables d_n, giving the set (d_n, T) of partial observables. If we do so, it may be convenient to express the correlations (d_n, T) as functions $d_n(T)$. A complete gauge invariant observable, fully predicted by the theory, is the value $d_n(T)$ of a planet distance at a certain given Earth proper time T from the initial event. Notice that T is not a coordinate. It is a complicated nonlocal function of the gravitational field, to which a measuring device (measuring a partial observable) has been attached. The use of a clock on Earth to determine a local temporal localization is just a matter of convenience.

Binary pulsar. Consider a binary system in which one of the two stars is a pulsar. Because of a Doppler effect, the frequency of the pulsing signal oscillates with the orbital period of the system. This fact allows us to count the number of pulses in each orbit. Let N_n be the number of pulses we receive in the *n-th* orbit. A theoretical model of the pulsar allows us to compute the expected decrease in orbital period due to gravitational wave emission and therefore the expected sequence N_n, which can be compared with the observed one. Doing this with sufficient care has won JH Taylor and RA Hulse the 1993 Nobel prize.

Notice that in all these examples the coordinates x^μ have disappeared from the observable quantities. This is true in general. A theoretical model of a physical system is made using coordinates x^μ, but then observable quantities are independent from the coordinates x^μ.[23]

2.3.2 The disappearance of spacetime

In the mathematical formalism of GR we utilize the "spacetime" manifold M, coordinatized by x. However, a state of the universe does not correspond to a configuration of fields on M. It corresponds to an equivalence class of field configurations under active diffeomorphisms. An active diffeomorphism changes the localization of the field on M by dragging it around. Therefore localization on M is just gauge: it is physically irrelevant.

In facts, M itself has no physical interpretation, it is just a mathematical device, a gauge artifact. Pre-general-relativistic coordinates x^μ design points of the physical spacetime manifold "where" things happen (see a detailed discussion below in 2.4.5); in GR there is nothing of the sort. M cannot be interpreted as a set of physical "events", or physical spacetime points "where" the

[23] Unless we gauge fix them to given partial observables. See 2.4.6.

fields take value. It is meaningless to ask whether or not the gravitational field is flat around the point A of M, because there is no physical entity "spacetime point A". Contrary to Newton and to Minkowski, there are no spacetime points where particle and fields live. There are no spacetime points at all. The Newtonian notions of space and time have disappeared.

In Einstein's words:

> " ...the requirement of general covariance takes away from space and time the last remnant of physical objectivity ..."[59]

Einstein justifies this conclusion in the immediate continuation of this text, which is the paragraph I have quoted at the end of the previous section, with the observation that all observations are space-time coincidences.

In Newtonian physics, if we take away the dynamical entities, what remains is space and time. In general relativistic physics, if we take away the dynamical entities, nothing remains. The space and time of Newton and Minkowski are reinterpreted as a configuration of one the fields, the gravitational field.

Concretely, this radically novel understanding of spatial and temporal relations is implemented in the theory by the invariance of the field equations under diffeomorphisms. Because of background independence –that is, since there are no non-dynamical objects that break this invariance in the theory– diffeomorphism invariance is formally equivalent to general covariance, namely the invariance of the field equations under arbitrary changes of the spacetime coordinates $\vec{x}$ and t.

Diffeomorphism invariance implies that the spacetime coordinates $\vec{x}$ and t used in GR have a different physical meaning than the coordinates $\vec{x}$ and t used in pre-relativistic physics. In pre-relativistic physics, $\vec{x}$ and t denote localization with respect to appropriately chosen reference objects. These reference objects are chosen in such a way that they make the physical influence of background spacetime manifest. In particular, their motion can be chosen to be inertial. In GR, on the other hand, the spacetime coordinates $\vec{x}$ and t have no physical meaning: physical predictions of GR are independent from the coordinates $\vec{x}$ and t.

A physical theory should not describe the location in space and the evolution in time of dynamical objects. It describes *relative* location and *relative* evolution of dynamical objects. Newton introduced the notion of background spacetime because he needed the acceleration of a particle to be well defined (so that $\vec{F} = m\vec{a}$ could make sense). In the newtonian theory and in special relativity, a particle accelerates when it does so with respect to a fixed spacetime in which the particle moves. In general relativity, a particle (a dynamical object) accelerates when it does so with respect to the local values of the gravitational field (another dynamical object). There is no meaning for the location of the gravitational field, or the location of the particle: only the relative location of the particle with respect to the gravitational field has physical meaning.

What remains of the prerelativistic notion of spacetime is a relation between dynamical entities: we can say that two particles' worldline "intersect"; that a field has a certain value "where" another field has a certain value; or that we measure two partial observables "together". This is precisely the modern realization of Descartes' notion of *contiguity*, and it is the basis of spatial and temporal notions in GR.

As Whitehead put it, we cannot have spacetime without dynamical entities, anymore than saying that we can have the cat's grin without the cat. The world is made by fields. Physically, these do not live on spacetime. They live, so to say, on one another. Not anymore fields on spacetime, just fields on fields. It is like in the metaphor in section 1.1.3 of the first chapter, where we had no more animals on the island, but just animals on the whale, animals on animals. Our feet are not anymore in space: we have to ride the whale.

2.4 * Complements

I close this chapter discussing a certain number of issues related to the interpretation of GR.

2.4.1 Mach principles

Ideas of Ernest Mach had a strong influence on Einstein's discovery of GR. Mach presents a number of acute criticisms to Newton's motivations for introducing absolute space and absolute time. In particular, he points out that in Newton's bucket argument there is a missing element: he observes that the inertial reference frame (the reference frame with respect to which rotation has detectable physical effects) is also the reference frame in which the fixed stars do not rotate. Mach suggests then that the inertial reference frame is not determined by absolute space, but rather it is determined by the entire matter content of the universe, including distant stars. He suggests that if we could repeat the experiment with a very massive bucket, the mass of the bucket would affect the inertial frame, and the inertial frame would rotate with the bucket.

In the light of GR, the observation is certainly pertinent and it is clear that the argument may have played a role in Einstein's dismissal of Newton's argument. However, for some reason, the precise relation between Mach's suggestion and GR has generated a vast debate. Mach's suggestion that inertia is determined by surrounding matter has been called "the Mach Principle" and much ink has been employed to discuss whether or not GR implements this principle; whether or not "GR is Machian". Remarkably, in the literature one finds arguments and proofs in favor as well as against the conclusion that GR is Machian. Why this confusion?

Because there is no well defined "Mach principle". Mach provided a very important but vague suggestion, that Einstein developed into a theory, not a precise statement that can be true or false. Every author that has discussed "the Mach Principle" has actually considered a *different* principle. Some of these "Mach principles" are implemented in GR, others are not.

In spite of the confusion, or perhaps thanks to it, the discussion on how machian is GR does shed some light on the physical content of GR. Here I list several versions of the Mach principle that have been considered in the literature, and, for each of these, I comment on whether this particular Mach principle is True or False in GR. In the following, "matter" means any dynamical quantity except the gravitational field.

- **Mach Principle 1:** *Distant stars can affect the local inertial frame.*
  True. Because matter affects the gravitational field.
- **Mach Principle 2:** *The local inertial frame is completely determined by the matter content of the universe.*
  False. The gravitational field has independent degrees of freedom.
- **Mach Principle 3:** *The rotation of the inertial reference frame inside the bucket is in fact dragged by the bucket, and this effect increases with the mass of the bucket.*
  True. In fact, this is the Lense-Thirring effect: a rotating mass drags the inertial frames in its vicinity.
- **Mach Principle 4:** *In the limit in which the mass of the bucket is large, the internal inertial reference frame rotates with the bucket.*
  Depends. It depends on the details of the way the limit is taken.
- **Mach Principle 5:** *There can be no global rotation of the universe.*
  False. Einstein believed this to be true in GR, but Goedel's solution is a counter-example.
- **Mach Principle 6:** *In the absence of matter, there would be no inertia.*
  False. There are vacuum solutions of the Einstein field equations.
- **Mach Principle 7:** *There is no absolute motion, only motion relative to something else, therefore the water in the bucket does not rotate in absolute terms, it rotates with respect to some dynamical physical entity.*
  True. This is the basic physical idea of GR.
- **Mach Principle 8:** *The local inertial frame is completely determined by the dynamical fields in the universe.*
  True. In fact, this is precisely Einstein key idea.

2.4.2 Relationalism versus substantivalism

In contemporary philosophy of science there is an interesting debate on the interpretation of GR. The two traditional thesis about space, absolute and relational, suitably edited to take into account scientific progress, continue under the names of *substantivalism* and *relationalism*. Here I present a few considerations on the issue.

GR changes the notion of spacetime in physics in the sense of relationalism. In prerelativistic physics, spacetime is a fixed nondynamical entity, over which physics happen. It is a sort of structured container which is the home of the world. In relativistic physics, there is nothing of the sort. There are only interacting fields and particles: the only notion of localization which is present in the theory is relative: dynamical objects can be localized only with respect to one another. This is the notion of space defended by Aristotle and Descartes, against which Newton wrote the initial part of the Principia. Newton had two points: the physical reality of inertial effects such as the concavity of

the water in the bucket, and the immense empirical success of his theory based on absolute space. Einstein provided an alternative interpretation for the cause of the concavity –interaction with the local gravitational field– and a theory based on relational space that has a better empirical success than Newton theory. After three centuries, the European culture has returned to a fully relational understanding of space and time.

At the basis of Cartesian relationalism is the notion of "contiguity". Two objects are contiguous if they are close to one another. Space is the order of things with respect to the contiguity relation. At the basis of the spacetime structure of GR is essentially the same notion. Einstein's "spacetime coincidences" are analogous to Descartes "contiguity".

A substantivalist position can nevertheless still be defended to some extent. Einstein's discovery is that Newtonian spacetime and the gravitational field, are the same entity. This can be expressed in two equivalent manners. One is that there is no spacetime: there is only the gravitational field. This is the choice I have made in this book. The second is that there is no gravitational field: it is spacetime that has dynamical properties. This choice is common in the literature. I prefer the first because I find that the differences between the gravitational field and the other fields are more accidental than essential. But the choice between the two points of view is only a matters of choice of words, and thus, ultimately, personal taste. If one prefers to keep the name "spacetime" for the gravitational field, then one can still hold a substantivalist position and claim that, according to GR, spacetime is an entity, not a relation. Furthermore, in GR localization can be defined with respect to the gravitational field, and therefore the substantivalist can say that spacetime is an entity that defines localization. For an articulation of this thesis, see for instance [84].

However, this is a very weakened substantivalist position. One is free to call "spacetime" anything with respect to which we define position. But to what extent is spacetime different from any arbitrary continuum of objects used to defined position? Newton's acute formulation of his substantivalism, already mentioned in a footnote above, contains a precise characterization of "space":

> "... So it is necessary that the definition of places, and hence of local motion, be referred to some motionless thing such as extension alone or "space", *in so far as space is seen to be truly distinct from moving bodies.*" [24]

The characterizing feature of space is that of being truly distinct from *moving* bodies, that is, in modern terms and after the Faraday-Maxwell conceptual revolution, that of being truly distinct from dynamical entities such as particles or fields. This is clearly not the case for the spacetime of GR. If the modern substantivalist is happy to give up Newton's strong substantivalism and identify the thesis that "spacetime is an entity" with the thesis that "spacetime is the gravitational field, which is a dynamical entity", then the distinction between substantivalism and relationalism is completely reduced to semantic.

When two opposite positions in a long standing debate have come so close that their distinction is reduced to semantic, one can probably say that the issue is solved. I think one can say that in this sense GR has solved the long standing issue of the relational versus substantivalist interpretation of space.

2.4.3 Has general covariance any physical content? Kretschmann's objection

Virtually any field theory can be reformulated in generally covariant form. An example of a generally covariant reformulations of a scalar field theory on Minkowski spacetime is presented below. This fact has lead some people to wonder whether general covariance has any physical significance at all. The argument is as follows: if any theory can be formulated in a general covariant language, then general covariance is not a principle that selects a particular class of theories, therefore it has no physical content. This argument was presented by Kretschmann shortly after Einstein's publication of GR. It is heard among some philosophers of science, and sometimes used also by some physicists that dismiss the conceptual novelty of GR.

I think that the argument is wrong. The non sequitur is the idea that a formal property that does not restrict the class of admissible theories, has no physical significance. Why should that be? Formalism is flexible, and we can artificially give a theory a certain formal property, especially if we accept byzantine formulations. But it does not follow from this that the use of one formalism or another is irrelevant. Physics is the search for the more effective formalism for reading Nature. The relevant question is not whether general covariance restricts the class of admissible theories. The relevant question is whether GR could have been conceived at all, or understood, without general covariance. Let me illustrate this point with the example of rotational invariance.

Kretschmann's objection applied to rotational symmetry. Ancient physics assumed that space has a preferred direction. The "up" and the "down" were considered absolutely defined. This changes with Newtonian physics, where space has rotational symmetry: all spatial directions are a priori equivalent, and only contingent circumstances –such as the presence of a nearby mass like the Earth– can make one direction particular. Physicists often say that

[24] I Newton, "De Gravitatione et aequipondio fluidorum", [56]

rotational invariance limits the admissible forces. But strictly speaking, this is not true. Kretschmann's objection applies equally well to rotational invariance: given a theory which is not rotationally invariant, we can reformulate it as a rotationally invariant theory, just by adding some variable. For instance, consider a physical theory T in which all bodies are subject to a force in the z direction $F = -g$, where g is a constant (such as gravity). This is a non-rotationally invariant theory. Now consider another theory T' in which there is a dynamical vector quantity $\vec{v}$, of length one, and a force $\vec{F} = g\vec{v}$. The theory T' is rotationally invariant, but in each solution the vector $\vec{v}$ will take a particular value, in a particular direction. Calling z this direction we have precisely the same phenomenology as theory T.

The example shows that we can express a non-rotationally-invariant theory T in a rotationally invariant formalism T'. Therefore rotational invariance does not *truly* restrict the class of admissible theories. Shall we conclude with Kretschmann that rotational invariance has no physical significance?

Obviously not. Modern physics *has* made a real progress with respect to ancient physics in understanding that space is rotationally invariant. Where is the progress? It is in the fact that the discovery of the rotational invariance of space puts us in far more effective position for understanding Nature. We can say that we have discovered that in general there is no preferred "up" and "down" in the universe. Equivalently, we can say that a rotationally invariant physical formalism is far more effective for understanding Nature than a non-rotationally invariant one.

The key points are two. First, it would have been difficult to find Newtonian theory within a conceptual framework in which the "up" and the "down" are considered absolute. Second, reformulating the theory T in the rotational invariant form T' modifies our understanding of it: we have to introduce the dynamical vector $\vec{v}$. From the point of view of the two theories T and T', the vector $\vec{v}$ is a byzantine construction without much sense. But notice that from the point of view of understanding Nature the introduction of $\vec{v}$ points to the physically correct direction: we are lead to investigate the nature and the dynamics of this vector. $\vec{v}$ is indeed the local gravitational field, and this is precisely the right track towards a more effective understanding of Nature. This is the strength of having understood rotational invariance.

In fact, if there is rotational invariance in the universe, there should be a rotationally invariant manner of understanding ancient physics, which, in its limited extent, was effective. Theory T' above represents precisely this better understanding of ancient physics. More than that, the reinterpretation itself indicates the new effective way of understanding the world. In conclusion, the fact that the effective but non-rotationally invariant theory T admits the byzantine rotationally invariant formulation T' is not an argument for the physical irrelevance of rotational invariance. Far from that, it is something that is required for us to have confidence in rotational invariance.

On the one hand, rotational invariance is interesting because it *enlarges*, not because it *restricts*, the kind of physics we can naturally describe. On the other hand, rotational invariance *does* drastically reduce the kind of theories that we are *willing* to consider. Not because it forbids us to write certain theories –such as theory T'–, but because if we want to describe a theory such as theory T we have to pay a price. Here the introduction of the vector $\vec{v}$. It is up to the theoretician to judge whether this price is worth paying, that is; whether $\vec{v}$ is in fact a physical entity worthwhile considering.

The value of a novel idea or a novel language in theoretical physics is not in the fact that old physics cannot be expressed in the new language. It is simply in the fact that it is more effective for describing reality. A physical theoretical framework is a map of reality. If the symbols of the map are better chosen, the map is more effective. A new language, by itself, rarely truly restricts the kind of theories that can be expressed. But it renders certain theories far simpler and others awkward. It orients our investigation on Nature. This, and nothing else, is scientific knowledge.

Let me come back to general covariance. Like rotational invariance, general covariance is a novel language, which expresses a general physical idea about the world. It is possible to express Newtonian physics in a generally covariant language. It is also possible to express GR physics in a non generally covariant language (by gauge fixing the coordinates). But Newtonian physics expressed in a covariant language or GR expressed in a noncovariant language are both monsters formulated in a form far more intricate than what it is possible. Nobody would have found them.

What Einstein has discovered is that two classes of entities previously considered distinct are in fact entities of the same kind. Newton taught us that (an effective way to understand the world is to think that) the world is made by two clearly *distinct* classes of entities, of very different nature. The first class is formed by space and time. The second class includes all dynamical entities moving in space and in time. In Newtonian physics these two classes of entities are different in many respect, and enter the formalism of physical models in very different manners. Einstein has understood that (a more effective way to understand the world is to think that) the world is *not* made by two distinct kinds of entities. There is only one type of entity: dynamical fields. General covariance is the language for describing a world without distinction between the spacetime entities and the dynamical entities. It is the language that does not assume this distinction.

We can reinterpret prerelativistic physics in a general covariant language. It suffices to rewrite the Newtonian absolute space and absolute time as a dynamical field, and then write generally covariant equations that fix them to their flat space values. But if we do so, we are not denying the physical content of Einstein's idea. On the contrary, we are simply reinterpreting the world in Einstein's terms. In other words, we are showing the strength, not the weakness of general covariance. Furthermore, in doing so we introduce a new physical field and we find ourself in the funny situation of having to write equations of motion for this field that constrain it to a single value. Thus we have

a theory where one of the dynamical fields is strangely constrained to a single value. This immediately suggests that perhaps we can relax a bit these equations and allow a full dynamics for this field. If we do so, we are directly on the track of GR. Again, far from showing the physical irrelevance of general covariance, this indicates its enormous cognitive strength.

I think that the mistake behind Kretschmann argument has two origins. The first is an excessively legalistic reading of the scientific enterprise. The second is the mistake of taking certain common physicists statements too literally. Physicists often write that a certain symmetry or a certain principle "uniquely determines" a certain theory. At a close reading, these statements are almost always far exaggerated. The uniqueness only holds under a vast number of other assumptions, that are left implicit, and which are facts or ideas the physicist considers natural, and does not bother detailing. The typical physicist carelessly dismisses counter examples by saying that they would be unphysical, implausible, or completely artificial. The connection between general physical ideas, general principles, intuitions, symmetries, is a burning melt of powerful ideas, not the icy demonstration of a mathematical theorem. What is at stake is finding the most effective language for thinking the world, not writing axioms. It is language in formation, not bureaucracy.[25]

General covariant flat space field theory. Consider the field theory of a free massless scalar field $\phi(x)$ on Minkowski space. The theory is defined by the action

$$S[\phi] = \int d^4x \; \eta^{\alpha\beta}\partial_\alpha\phi \; \partial_\beta\phi. \tag{2.138}$$

The equation of motion is the flat space Klein-Gordon equation

$$\eta^{\alpha\beta}\partial_\alpha\partial_\beta\phi = 0 \tag{2.139}$$

and the theory is obviously not generally covariant.

A trivial way to reformulate this theory in a general covariant language is to introduce the tetrad field $e^\alpha_\mu(x)$ and write the equations

$$\partial_\mu(e\,\eta^{\alpha\beta}e^\mu_\alpha e^\nu_\beta\partial_\nu\phi) = 0, \tag{2.140}$$

$$R^\alpha{}_{\beta\mu\nu} = 0. \tag{2.141}$$

The solution of (2.141) is that e is flat. Since the system is covariant we can chose a gauge in which $e^\alpha_\mu(x) = \delta^\alpha_\mu$. In this gauge, (2.140) becomes (2.138).

A more interesting way, is as follows. Consider a field theory for five scalar fields $\Phi^A(x)$, where $A = 1,\dots,5$. Use the notation

$$V_A = \epsilon_{ABCDE}\;\partial_\mu\Phi^B\partial_\nu\Phi^C\partial_\rho\Phi^D\partial_\sigma\Phi^E\;\epsilon^{\mu\nu\rho\sigma}, \tag{2.142}$$

where $\epsilon^{\mu\nu\rho\sigma}$ and ϵ_{ABCDE} are the 4-dimensional and 5-dimensional completely antisymmetric pseudotensors. Consider the theory defined by the action

$$S[\Phi^A] \;=\; \int d^4x \;\; V_5^{-1}(V_4V_4 - V_3V_3 - V_2V_2 - V_1V_1) \tag{2.143}$$

where V_5 is assumed never to vanish. The theory is invariant under diffeomorphisms. Indeed, V_A transforms as a scalar density (because $\epsilon^{\mu\nu\rho\sigma}$ is a scalar density), hence the integrand is a scalar density and the integral is invariant. For $\alpha = 1,2,3,4$, define the matrix

$$E^\alpha_\mu(x) = \partial_\mu\Phi^\alpha(x), \tag{2.144}$$

its inverse E^μ_α and its determinant E. Varying Φ^5, we obtain the equation of motion

$$\partial_\mu(E\;\eta^{\alpha\beta}E^\mu_\alpha E^\nu_\beta\partial_\nu\Phi^5) = 0. \tag{2.145}$$

This is the massless Klein-Gordon equation (2.140) interacting with a gravitational field E^α_μ. Varying Φ^α we do not obtain independent equations. We obtain the energy momentum conservation law implied by (2.145). The fact that there is only one independent equation is a consequence of the fact that there is a four-fold gauge invariance. We can chose a gauge in which

$$\Phi^a(x) = x^a \tag{2.146}$$

---

[25] Historically, the entire issue might be the result of a misunderstanding. Kretschmann attacked Einstein in a virulent form. In particular, he attacked Einstein's coincidence's solution of the hole argument. Now, Einstein probably learnt this idea, namely the idea that coincidences are the only observables, precisely from Kretschmann, but didn't give much credit to Kretschmann for this. I suppose this should have made Kretschmann quite bitter. I think that Kretschmann's subtext in saying that general covariance is empty was not that general covariance was no progress with respect to old physics: it was that general covariance was no progress with respect to what he himself had already realized before Einstein.

We have then immediately $E^a_\mu = \delta^a_\mu$, and equation (2.145) becomes (2.138). The other four equations are

$$\partial_a(\partial^a \Phi^5 \partial_b \Phi^5 - 1/2\delta^a_b \partial_c \Phi^5 \partial_c \Phi^5) = 0. \tag{2.147}$$

Even better, we may not fix the gauge, and consider the gauge invariant function of four variables $\phi(X^a)$ defined by

$$\phi(\Phi^a(x)) \quad = \quad \Phi^5(x). \tag{2.148}$$

This function satisfies the Minkowski space Klein-Gordon equation (2.138).

How to interpret such a theory? The theory (2.138) is not generally covariant, therefore its coordinates x are (partial) observables. The theory is defined by five partial observables: four x^μ and ϕ. To interpret the theory we must have measuring procedures associated to these five quantities. The relation between between these observables is governed by equation (2.138). On the other hand, the theory (2.143) is generally covariant; therefore the coordinates x are not observable. The theory is defined by five partial observables: the five ϕ^A. We must have measuring procedures associated to these five quantities. The relation between between them observables is governed by again by equation (2.138). Therefore in the two cases we have the same partial observables, identified by $\Phi^A \leftrightarrow (x^a, \phi)$, related by the same equation.

There is only one subtle but important difference between the theory (2.143) and theory (2.138): the theory (2.138), separates the five partial observables (x, ϕ) into two sets: the independent ones (x) and the dependent one (ϕ). On the other hand, the theory (2.143) treats the five partial observables Φ^A on the equal footing. Thus, in a strict sense, the theory (2.138) contains one extra information: a distinction between dependent and independent partial observables. Because of this difference, the two theories reflect two quite different interpretations of the world. The first describes a world's ontology split into spacetime and matter. The second describes a world where spatiotemporal relation are interpreted as relational.

2.4.4 Meanings of time

The concept of time used in natural language carries many properties. Within a given theoretical framework (say Newtonian mechanics) time maintains some of these properties, and loses others. In different theoretical frameworks time has different properties. The best known example is probably the directionality of time: absent in mechanics, present in thermodynamics. But there are many other features of time that are lacking in one theory and are present in others. For instance, a property of time in Newtonian mechanics is uniqueness: there is a unique time interval between any two events; on the contrary, in special relativity there are as many time variables as there are Lorentz observers $(x^0, x'^0 \ldots)$. Another attribute of time in Newtonian mechanics is globality: every solution of the equations of motion "passes" through every value of Newtonian time t once and only once. In some cosmological models, on the other hand, there is no choice of time variable with such a property: there is "no time", if we demand that being global is an essential property of time. In other words, we use the word "time" to denote quite different concepts, that may or may not include this or that property.

Here I describe a simple classification of possible attributes of time. I identify ten separate levels of increasing complexity of the notion of time, corresponding to an increasing number of properties. Theories fall in one or the other of these levels, according to the set of attributes that the theory ascribe to the notion of time it uses. The tenfold arrangement is conventional: the main point I intend to emphasize is that a single, clear and pure notion of "time" does not exist.

Properties of time. Consider an infinite set S without any structure. Add to S a topology and a differential structure dx. S becomes a manifold; assume that this manifold is one dimensional, and denote the set S together with its differentiable structure as the line $L = (S, dx)$. Next, assume we add a metric structure d to L; denote the resulting metric line as $M = (S, dx, d)$. Next, fix an ordering $<$ (a direction) in M. Denote the resulting oriented line as the affine line $A = (S, dx, d, <)$. Next, fix a preferred point of A as the origin 0; the resulting space is isomorphic to the real line $R = (S, dx, d, <, 0)$.

The real line R is the traditional metaphor for the idea of time. Time is frequently represented by a variable t in R. The structure of R corresponds to an ensemble of properties that we naturally associate to the notion of time. These are the following. (a) The existence of a topology on the set of the time instants, namely the existence of a notion of two time instants being close to each other, and the fact that time is "one dimensional". (b) The existence of a metric. Namely the possibility of stating that two distinct time intervals are equal in magnitude. Time is "metric". (c) The existence of an ordering relation between time instants. Namely, the possibility of distinguishing the past direction from the future direction. (d) The existence of a preferred time instant, the present, the "now". To capture these properties in mathematical language, we describe time as a real line R. An affine line A describes time up to the notion of present; a metric line M describes time up to the notions of present and past/future distinction; a line L describes time up to the notion of metricity.

In Newtonian mechanics, we begin by representing time as a variable in R, but then the equations are invariant both under $t \mapsto -t$ and under $t \mapsto t + a$. Thus the theory is actually defined in terms of a variable t in a metric line M. Newtonian mechanics, in fact, incorporates both the notions of topology of the set of time instants and (in a very essential way) the fact that time is metric, but it does not make any use of the notion of present, nor the

direction of time. This is well known. Note that Newtonian theory is not inconsistent with the introduction of the notions of a present and of time-directionality: it simply does not make any use of these notions. These notions are not present in Newton theory.

The properties listed above do not exhaust the different ways in which the notion of time enters physical theories; the development of theoretical physics has modified substantially the natural notion of time. A first modification was introduced by special relativity. Einstein's definition of the time coordinate of distant events yields a notion of time which is observer dependent. An invariant structure can be maintained at the price of relaxing the 1d character of time and the 3d character of space, in favor of a notion of 4d spacetime. Alternatively, we may say that the notion of a single time is replaced by a three parameter family of times $t_{\vec{v}}$, one for each Lorentz observer. Therefore, the time we use in special relativity is not unique as the time of Newtonian mechanics. Rather than a single line, we have a three parameter family of lines (the straight lines through the origin that fill the light cone of Minkowski space). Denote this three-parameters set of lines as M^3.

Times in GR. There are several distinct possibilities of identifying "time" in GR. Each singles out a different notion of time. Each of these notions reduce to the standard non-relativistic or special relativistic time in appropriate limits, but each lacks at least some of the properties of nonrelativistic time. The most common ways of identifying time within GR are the following.

Coordinate time x^0. Coordinate time can be arbitrarily rescaled, and does not provide a way to identifying two time intervals as equal in duration. Therefore it is not metric, in the sense defined above. In addition, the possibility of changing time coordinate freely from point to point implies that there is an infinite dimensional choice of equally good coordinate times. Finally, unlike prerelativistic time, x^0 is not an observable quantity. Denote the set of all the possible coordinate times as L^∞.

Proper time τ. This notion of time is metric. But it is very different from the notion of time in special relativity for several reasons. First, it is determined by the gravitational field. Second, we have a different time for each worldline, or, infinitesimally, for every speed at every point. For an infinitesimal timelike displacement dx^μ at a point x, the infinitesimal time interval is $d\tau = \sqrt{g_{\mu\nu}(x)\, dx^\mu dx^\nu}$. This notion of time is a radical departure from the notion of time used in special relativity because it is determined by the dynamical fields in the theory. A solution of Einstein equations defines a point in the phase space Γ of GR. It assigns a metric structure to every world line. Therefore this notion of time is given by a function from the phase space Γ times the set of the world lines wl into the metric structures $d : wl \times wl \mapsto R^+$. Denote this function as m^∞. Call "internal" a notion of time affected by the dynamics.

Before GR, dynamics can be expressed as evolution in a single time variable which has metric properties and can be measured. In general relativistic physics, this concept of time splits into two distinct concepts: we can still view the dynamics as evolution in a time variable -x^0-, but this time has no metric properties and is not observable; alternatively, there is a notion of time that has metric properties -τ-, but the dynamics of the theory cannot be expressed as evolution in τ. Is there a way to go around this split and view GR as a dynamical theory in the sense of a theory expressing evolution in an observable metric time?

Clock time. The dynamics of GR determines how observable quantities evolve with respect to one another. We can always choose one observable quantity t_c, declare it the independent one, and describe how the other observables evolve as functions of it. A typical example of this clock time is the radius of a spatially compact universe in relativistic cosmology R. Formally, clock time is a function on the extended configuration space $\mathcal{C}$ of the theory (see Chapter 3.) Denote this notion of time as the clock time $c : \mathcal{C} \mapsto R$.

Under this definition of time, GR becomes similar to a standard hamiltonian dynamical theory. A clock time, however, generally behaves as a clock only in certain states or for a limited amount of time. The radius of the universe, for instance, fails as a good time variable when the universe recollapses. In general a clock time lacks temporal globality. In fact, several results are known [92], concerning obstructions to defining a function t_c that behaves as "a good time" globally.

Notice that some of these relativistic notions of time are, in a sense, opposite to the prerelativistic case: while in Newtonian theory time evolution is captured by a function from the metric line M (time) to the configuration or phase space, now the notion of time is captured by a function from the configuration or phase space to the metric line. This inversion is the mathematical expression of the physical idea that the flow of time is affected or determined by the dynamics of the system itself.

Finally, none of the ways in which time can be thought in classical GR can be uncritically extended to the quantum regime. Quantum fluctuations of physical clocks, and quantum superposition of different metric structures make the very notion of time fuzzy at the Planck length. As will be discussed in the second part of the book, a fundamental concept of time may be absent in quantum gravity.

Times. Notice that properties of time progressively disappear in going toward more fundamental physical theories. At the opposite end of the spectrum, there are properties of time associated with the notion of time used in the natural languages, which are not present in physical theories. They play a role in other areas of natural investigations.

I mention these properties for the sake of completeness. These are for instance memory, expectations, and the psychological perception of free will.

To summarize, I have identified the following properties of the notion of time.

1. Existence of memory and expectations.
2. Existence of a preferred instant of time, the present, the now.
3. Directionality: the possibility of distinguishing the past from the future direction.
4. Uniqueness: the feature that is lost in special and general relativity, where we cannot identify a preferred time variable.
5. The property of being external: the independence of the notion of time from the dynamical variables of the theory.
6. Spatial globality: the possibility of defining the same time variable in all space points.
7. Temporal globality: the fact that every motion goes through every value of the time variable once and only once.
8. Metricity: the possibility of saying that two time intervals have equal duration
9. One dimensionality, namely the possibility of arranging the time instants in a one dimensional manifold.

This discussion suggests a sequence notions of time, which I list here in order of decreasing complexity.

Time of natural language. This is the notion of time of everyday language, which includes all the features listed in the previous section. This notion of time is not necessarily non-scientific: for instance, any scientific approach to, say, the human brain, should make use of this notion of time.

Time-with-a-present. This is the notion of time that have all the features listed in the previous section, including the existence of a preferred instant, the present, but up to the notions of memory and expectations, which are notions usually related more to complex systems (brain) than to time itself. The notion of present is generally considered a feature of time itself. This notion of time is the one to which often people refer when they refer to the "flow of time" or Eddington's "vivid perception of the flow of time" [60]. This notion of time can be described by the structure of a parametrized line R.

Thermodynamical time. If we maintain the distinction between a future direction and a past direction, but we give up the notion of present, we obtain the notion of time typical of thermodynamics. Since thermodynamics is the first physical science that appear in this list, this is maybe a good point to emphasize that the notion of present, of the "now", is completely absent from the description of the world in physical terms. This notion of time can be described by the structure of an affine line A.

Newtonian time. In Newtonian mechanics there is no preferred direction of time. Notice that in the absence of a preferred direction of time the notions of cause and effect are interchangeable. This notion of time can be described by the structure of a metric line M.

Special relativistic time. If we give up uniqueness, we have the time used in special relativity: different Lorentz observers have a different notion of time. Special relativistic time is still external, spatially and temporally global, metrical and one dimensional, but it is not unique: There is a three parameters set of quantities that share the status of time. This notion of time can be described by the three parameters set of metric lines M^3.

Cosmological time. By this I indicate a time which is spatially and temporally global, metrical and one dimensional, but it is not external, namely it is dynamically determined by the theory. Proper time in cosmology is the typical example. It is the most structured notion of time that occurs in GR. Denote it by m.

Proper time. By this I indicate a time which is temporally global, metrical and one dimensional, but it is not spatially global, as the notion of proper time along world lines in GR. It can be represented by a function m^∞ defined on the Cartesian product of the phase space and the ensemble of the world lines.

Clock time. By this I indicate a time which is metrical and one dimensional, but it is not temporally global. A realistic matter clock in GR defines a time in this sense. This notion of time can be described by a function c on the phase space.

Parameter time. By this we mean a notion of time which is not metric and not observable. The typical example is the coordinate-time in GR. Another example of parameter time is the evolution parameter in the parametrized formulation of the dynamics of a relativistic particle. Parameter time is described by an unparametrized line L, or by an infinite set L^∞ of unparametrized lines.

No-time. Finally, this is the bottom level in the analysis; it is not a time concept, but rather I indicate by no-time the idea about time underlying every theory in which there is no fundamental notion of time at all.

The list must of course not be taken rigidly. It is summarized in the Table 2.1.

There is a interesting feature that emerges from the above analysis: the hierarchical arrangement. While some details of this arrangement may be artificial, nevertheless the analysis points to a general fact: moving from theories of "special" objects, like the brain or the living beings, toward more general theories that include larger portions of Nature, we make use of a physical notion of time that is less specific and has less determinations. If we observe Nature at progressively more fundamental levels, and we seek for laws that hold in more general contexts, then we discover that these laws require or admit an increasingly weaker notion of time.

Table 2.1: Times.

Time notion	Property	Example	Form
Natural language time	memory	brain	?
Time-with-a-present	present	biology	R
Thermodynamical time	direction	thermodynamics	A
Newtonian time	unique	newtonian mechanics	M
Special relativistic time	external	special relativity	M^3
Cosmological time	spatially global	cosmological time	m
Proper time	temporally global	world line proper time	m^∞
Clock time	metric	clocks in GR	c
Parameter time	one dimensional	coordinate time	L^∞
No-time	none	quantum gravity	none

This observation suggests that "high level" features of time are not present at the fundamental level, but "emerge" as features of specific physical regimes, like the notion of "water surface" emerges in certain regimes of the dynamics of a combination of water and air molecules (see for instance [61]).

Notions of time with more attributes are high level notions that have no meaning in more general situations. The uniqueness of Newtonian time for instance, make sense only in the special regime in which we consider an ensemble of bodies moving slowly with respect to each other. Thus, the notion of a unique time is high level notion that makes sense only for some regimes in Nature. For general systems, most features of time are genuinely meaningless.

2.4.5 Nonrelativistic coordinates

The precise meaning of the coordinates $x = (\vec{x}, t)$ in Newtonian and special relativistic physics is far from obvious. Let me recall it here, in order to clarify the precise difference with the relativistic coordinates.

Newton is well aware that the motions we observe are relative motions, and stresses this point in the Principia. His point is not that we can directly *observe* absolute motion. His point is that we can *infer* the absolute motion or "true motions", or motion with respect to absolute space, from its physical effects (such as the concavity of the water in the bucket), starting from our observation of relative motions.

For instance, we observe and describe motions with respect to Earth; but from subtle effects, such as Foucault's pendulum, we infer that these are not true motions. The experiment of the bucket is an example of the possibility of revealing true motion (rotation of the water with respect to space), disentangling it from relative motion (rotation with respect to the bucket), by means of an observable effect (the concavity of the water surface).[26]

For Newton, the coordinates $\vec{x}$ that enter his main equation

$$\vec{F} = m \frac{d^2\vec{x}(t)}{dt^2}. \tag{2.149}$$

are the coordinates of absolute space. However, since we cannot directly observe space, the only way we have to coordinatize space points is by using physical objects. The coordinates $\vec{x}$ of the object A moving along the trajectory $\vec{x}(t)$ are therefore defined as distances from a chosen system O of objects, which we call a "reference frame". But then $\vec{x}$ are not the coordinates of absolute space. So, how can equation (2.149) work?

The solution of the difficulty, is to use the capacity of unveiling "true motion" that Newton has pointed out, in order to select the objects forming the reference frame O wisely. There are "good" and "bad" reference frames. The good ones are the ones in which no effect such as the concavity of the water surface of Newton's bucket can be observed, within a desired accuracy. Equation (2.149) is correct, to the desired accuracy, if we use coordinates defined with respect to these good frames. In other words, the physical content of (2.149) is actually quite subtle:

[26] Newton accords deep significance to the fact that we can unveil true motion. He describes relative motion as the way reality is observed by us, and true motion as the way reality might be directly "perceived", or "sensed", by God. This is why Newton calls space –the entity with respect to which true motion happens– the "Sensorium of God": true motion is motion "with respect to God", or "as perceived by God". There is a platonic tone in this idea that reason finds the way to the veiled divine truth beyond appearances. I wouldn't read this as so removed from modernity as it is often portrayed. There isn't all that much difference between Newton's inquiry in God's way of "sensing the world", and the modern search for the most effective way of conceptualizing reality. Newton's God plays a mere linguistical role here: the role of denoting a major enterprise: upgrading our own conceptual structure for understanding reality.

> There exist reference objects O with respect to which the motion of any other object A is correctly described by (2.149).

This is a statement that begins to be meaningful only when a sufficiently large number of moving objects is involved.

Notice also that for this construction to work it is important that the objects O forming the reference frame are not affected by the motion of the object A. There shouldn't be any dynamical interaction between A and O.

Special relativity does not change much of this picture. Since absolute simultaneity makes no sense, if the event A is distant from the clock in the origin, its time t is ill defined. Einstein's idea is to *define* a procedure for assigning a t to distant events, using clocks moving inertially.

> At clock time t_e, send a light signal that reaches the event. Receive the reflected signal back at t_r. The time coordinate of the event is *defined* to be $t_A = \frac{1}{2}(t_e + t_r)$.

It is important to emphasize that this is a useful definition, not a metaphysical statement that the event A happens *"right at the time when"* the observer clock displays t_A.

Special relativity replaces Newton's absolute space and absolute time with a single entity: Minkowski's absolute spacetime, while the notion of inertial system and the meaning of the coordinates are the same as in Newtonian mechanics.

Summarizing, these coordinates have the following properties.

(a) Coordinates describe position with respect to physical reference objects (reference frames).

(b) Space coordinates are defined by the *distance* from the reference bodies. Time coordinates are defined with respect to isochronous clocks.

(c) Reference objects are appropriately chosen: they are such that the reference system they define is inertial.

(d) Inertial frames reveal the structure of absolute spacetime itself.

(e) The objects A whose dynamics is described by the coordinates do not interact with the reference objects O. There is no dynamical coupling between A and O.

Relativistic coordinates do not have *any* of these properties. The fact that the two are indicated with the same notation x^μ is only un unfortunate historical accident.

2.4.6 Physical coordinates and GPS observables

Instead of working with arbitrary unphysical coordinates x^μ, we can choose to coordinatize spacetime events with coordinates X^μ having an assigned physical interpretation. For instance, we can describe the universe by giving a name $\vec{X}$ to each galaxy, and choosing X^0 as the proper time from the big bang, along the galaxy world line. If we do so, the defining properties of the coordinates X must be added to the formalism. We must add a certain number of equations on the gravitational field: the equations of motions of the objects used to fix the coordinates (the galaxies, in the example). These additional equations gauge fix general covariance.

The gauge fixing can also be partial. For instance, a common choice is

$$\hat{e}^0_0(X) = 1, \qquad \hat{e}^i_0(X) = 0, \qquad \hat{e}^0_a(X) = 0. \tag{2.150}$$

where $i = 1, 2, 3$ and $a = 1, 2, 3$. This correspond to partially fixing the coordinates by requiring that X^0 measures proper time, that equal X^0 surfaces are locally instantaneity surfaces, in the sense of Einstein, for the constant $\vec{X}$ lines and that the local Lorentz frames are chosen so that these lines are still.

If the coordinates are fully specified, the set formed by these physical gauge fixing equations and the equations of motion has no residual gauge invariance; that is, initial data determine evolution uniquely. This procedure can be implemented in many possible ways, since there are arbitrarily many ways of fixing physical coordinates, and none is a priori better than the others. In spite of this arbitrariness, this procedure is often convenient, particularly when the particular physical situation suggests a natural coordinate choice, as in the cosmological context mentioned.

Physical coordinates X^μ defined by matter filling up space can only be effectively used in the cosmological context because it is only at the cosmological scale that matter fills up space. In a system in which there are empty regions, such as the solar system, these physical coordinates are not available. An interesting alternative choice is provided by the GPS coordinates described below.

The physical coordinates X^μ are partial observables and we can associate measuring devices to them.

Undetermined physical coordinates. Finally, there is still a third interpretation of the coordinates of GR, which is intermediate between arbitrary coordinates x^μ and physical coordinates X^μ. Imagine that a region of the universe is filled with certain light objects, which may not be in free fall. We can use these objects to define physical coordinates X^μ, but also choose to ignore the equations of motion of these objects. We obtain a system of equation for the gravitational field and other matter, expressed in terms of coordinates X^μ that are interpreted as the spacetime location of reference objects whose dynamics we *have chosen* to ignore.

This set of equation is underdetermined: same initial conditions can evolve into different solutions. However, the interpretation of such underdetermination is simply that we have chosen to neglect part of the equations of motion. Different solutions with the same initial conditions represent the same physical configuration of the fields,

but expressed, say, in one case with respect to free falling reference objects, in the other case with respect to reference objects on which a force has acted at a certain moment, and so on. This procedure has the disadvantage of being useless in quantum theory, where we cannot assume that something is observable and at the same time neglect its dynamics.

In conclusion, one should always be carefully, in talking about general relativistic coordinates, whether one is referring to

(i) arbitrary mathematical coordinates x,

(ii) physical coordinates X with an interpretation as positions with respect to objects whose equations of motions are taken into account,

(iii) physical coordinates with an interpretation as positions with respect to objects whose equations of motions are ignored.

The system of equations of motion is non deterministic in (i) and (iii), deterministic in (ii). The coordinates are partial observables in (ii) and (iii), but not in (i). Confusion about observability in GR follows from confusing these three different interpretations of the coordinates. I describe in the following an example of physical coordinates.

GPS observables. In the literature there are many attempts to define useful physical coordinates. It is easier to define physical coordinates in the presence of matter than in the context of pure GR. Ideally, we can consider GR interacting with four scalar matter fields. Assume that the configuration of these fields is sufficiently nondegenerate. Then the components of gravitational field at points defined by given values of the matter fields are gauge invariant observables. This idea has been developed in a number of variants, such as dust carrying clocks and others (see [62, 85, 63] and references therein). The extent to which the result is realistic or useful is questionable. It is rather unsatisfactory to understand the theory in terms of fields that do not exist, or phenomenological objects such as dust, and it is questionable whether these procedures could make sense in the quantum theory, where the aim is to describe Planck scale. Earlier attempts to write a complete set of gauge invariant observables are in the context of pure GR [64]. The idea is to construct four scalar functions of the gravitational field (say, scalar polynomials of the curvature), and use these to localize points. The value of a fifth scalar function in a point where the four scalar functions have a given value is a gauge invariant observable. This works, but the result is mathematically very intricate and physically very unrealistic. It is certainly possible, in principle, to construct detectors of such observables, but I doubt any experimenter would get funded for a proposal to build such apparatus.

There is a simple way out, based on GR coupled with a minimal and *very* realistic amount of additional matter. Indeed, this way out is so realistic that it is in fact real: it is essentially already implemented by existing technology, the Global Positioning System (GPS), the first technological application of GR, or the first large scale technology that needs to take GR effects into account [65].

Consider a general covariant system formed by GR coupled with four small bodies. These are taken to have negligible mass; they will be considered as point particles for simplicity, and called "satellites". Assume that the four satellites follow timelike geodesics; that these geodesics meet in a common (starting) point O; and, that at O they have a given (fixed) speed –the same for the four– and directions as the four vertices of a tetrahedron. The theory might include any other matter. Then (there is a region $\mathcal{R}$ of spacetime for which) we can uniquely associate four numbers $s^\alpha, \alpha = 1, 2, 3, 4$ to each spacetime point p as follows. Consider the past lightcone of p. This will (generically) intersect the four geodesics in four points p_α. The numbers s^α are defined as the distance between p_α and O. (That is: the proper time along the satellites' geodesic.) We can use the s^α's as physically defined coordinates for p. The components $g_{\alpha\beta}(s)$ of the metric tensor in these coordinates are gauge invariant quantities. They are invariant under four-dimensional diffeomorphisms (because these deform the metric as well as the satellites' worldlines). They define a complete set of gauge invariant observables for the region $\mathcal{R}$.

The physical picture is simple, and its realism is transparent. Imagine that the four "satellites" are in fact satellites, each carrying a clock that measures the proper time along its trajectory, starting at the meeting point O. Imagine also that each satellite broadcasts its local time with a radio signal. Suppose I am at the point p and have an electronic device that simply receives the four signals and displays the four readings. See Figure 1. These four numbers are precisely the four physical coordinates s^α defined above. Current technology permits to perform these measurements with accuracy well within the relativistic regime [65, 66]. If I then use a rod, and a clock and measure the physical distances between s^α coordinates points, I am directly measuring the components of the metric tensor in the physical coordinate system. In the terminology of chapter 3, the s^α's are *partial* observables, while $g_{\alpha\beta}(s)$ are *complete* observables.

As shown below, the physical coordinates s^α have nice geometrical properties; they are characterized by

$$g^{\alpha\alpha}(s) = 0, \qquad \alpha = 1, \dots, 4. \tag{2.151}$$

Surprisingly, in spite of the fact that they are defined by what looks like a rather nonlocal procedure, the evolution equations for $g_{\alpha\beta}(s)$ are local. These evolution equations can be written explicitly using the Arnowitt-Deser-Misner (ADM) variables [120]. Lapse and Shift turn out to be fixed local functions of the three metric.

In what follow, I first introduce the GPS coordinates s^α in Minkowski space. Then I go over to a general spacetime. I assume the Einstein summation convention only for couples of repeated indices that are one up and one

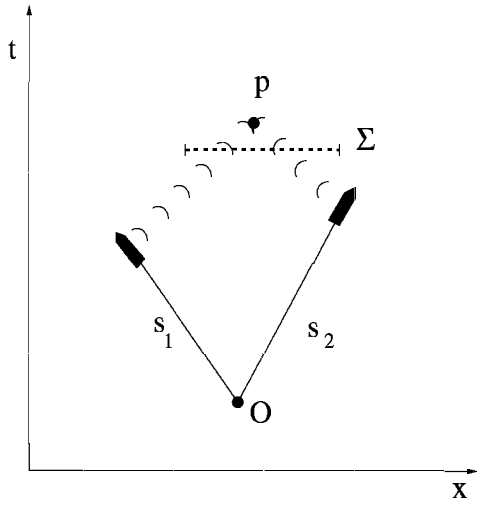


Figure 2.5: s_1 and s_2 are the GPS coordinates of the point p. Σ is a Cauchy surface with p in its future domain of dependence.

down. Thus α is not summed over in (2.151). While dealing with Minkowski spacetime, the spacetime indices μ, ν are raised and lowered with the Minkowski metric. Here I write here an arrow over three- as well as four-dimensional vectors.

Consider a tetrahedron in three-dimensional euclidean space. Let its center be at the origin and it four vertices $\vec{v}^\alpha$, where and $\vec{v}^\alpha \cdot \vec{v}^\beta = -1/3$ for $\alpha \neq \beta$, have unit length $|\vec{v}^\alpha|^2 = 1$. Here $\alpha = 1, 2, 3, 4$ is an index that distinguishes the four vertices, and should not be confused with vector indices. With a convenient orientation, these vertices have cartesian coordinates ($a = 1, 2, 3$)

$$v^{1a} = (0,\ 0,\ 1), \qquad v^{2a} = (2\sqrt{2}/3,\ 0,\ -1/3), \tag{2.152}$$

$$v^{3a} = (-\sqrt{2}/3,\ \sqrt{2/3},\ -1/3), \qquad v^{4a} = (-\sqrt{2}/3,\ -\sqrt{2/3},\ -1/3). \tag{2.153}$$

Let us now go to a four dimensional Minkowski space. Consider four timelike 4-vectors $\vec{W}^\alpha$, of length one, $|\vec{W}^\alpha|^2 = 1$, representing the normalized 4-velocities of four particles moving away from the origin in the directions $\vec{v}^\alpha$ at a common speed v. Their Minkowski coordinates ($\mu = 0, 1, 2, 3$) are

$$W^{\alpha\mu} = \frac{1}{\sqrt{1-v^2}}\,(1,\ v\,v^{\alpha a}). \tag{2.154}$$

Fix the velocity v by requiring the determinant of the matrix $W^{\alpha\mu}$ to be one one. (This choice fixes v at about half the speed of light; a different choice changes only a few normalization factors in what follows.) The four by four matrix $W^{\alpha\mu}$ plays an important role in what follows. Notice that it is a fixed matrix whose entries are certain given numbers.

Consider one of the four 4-vectors, say $\vec{W} = \vec{W}^1$. Consider a free particle in Minkowski space that starts from the origin with 4-velocity $\vec{W}$. Call it a "satellite". Its world line l is $\vec{x}(s) = s\vec{W}$. Since $\vec{W}$ is normalized, s is precisely the proper time along the world line. Consider now an arbitrary point p in Minkowski spacetime, with coordinates $\vec{X}$. Compute the value of s at the intersection between l and the past light cone of p. This is a simple exercise, giving,

$$s = \vec{X} \cdot \vec{W} - \sqrt{(\vec{X} \cdot \vec{W})^2 - |\vec{X}|^2}. \tag{2.155}$$

Now consider four satellites, moving out of the origin at 4-velocity $\vec{W}^\alpha$. If they radio broadcast their position, an observer at the point p with Minkowski coordinates $\vec{X}$ receives the four signals s^α

$$s^\alpha = \vec{X} \cdot \vec{W}^\alpha - \sqrt{(\vec{X} \cdot \vec{W}^\alpha)^2 - |\vec{X}|^2}. \tag{2.156}$$

Introduce (non-Lorentzian) general coordinates s^α on Minkowski space, defined by the change of variables (2.156). These are the coordinates read out by a GPS device in Minkowski space. The Jacobian matrix of the change of

coordinates is given by

$$\frac{\partial s^\alpha}{\partial x^\mu} = W^\alpha_\mu - \frac{W^\alpha_\mu(\vec{X}\cdot\vec{W}^\alpha) - X_\mu}{\sqrt{(\vec{X}\cdot\vec{W}^\alpha)^2 - |\vec{X}|^2}}, \tag{2.157}$$

where W^α_μ and X_μ are $W^{\alpha\mu}$ and X^μ with the spacetime index lowered with the Minkowski metric. This defines the the tetrad field $e^\alpha_\mu(s)$

$$e^\alpha_\mu(s(X)) = \frac{\partial s^\alpha}{\partial x^\mu}(X). \tag{2.158}$$

The contravariant metric tensor is given by $g^{\alpha\beta}(s) = e^\alpha_\mu(s)e^{\mu\beta}(s)$. Using the relation $|\vec{W}^\alpha|^2 = 1$, a straightforward calculation shows that

$$g^{\alpha\alpha}(s) = 0, \qquad \alpha = 1,\dots,4. \tag{2.159}$$

This equation has the following nice geometrical interpretation. Fix α and consider the one-form field $\omega^\alpha = ds^\alpha$. In s^α coordinates, this one-form has components $\omega^\alpha_\beta = \delta^\alpha_\beta$, and therefore "length" $|\omega^\alpha|^2 = g^{\beta\gamma}\omega^\alpha_\beta\omega^\alpha_\gamma = g^{\alpha\alpha}$. But the "length" of a 1-form is proportional to the volume of the (infinitesimal, now) 3-surface defined by the form. The 3-surface defined by ds^α is the surface $s^\alpha = constant$. But $s^\alpha = constant$ is the set of points that read the GPS coordinate s^α, namely that receive a radio broadcasting from a same event p_α of the satellite α, namely that are on the future light cone of p_α. Therefore $s^\alpha = constant$ is a portion of this light cone, therefore it is a null surface, therefore its volume is zero, therefore $|\omega^\alpha|^2 = 0$, therefore $g^{\alpha\alpha} = 0$.

Since the s^α coordinates define $s^\alpha = constant$ surfaces that are null, we denote them as "null GPS coordinates". It is useful to introduce another set of GPS coordinates as well, which have the traditional timelike and spacelike character. We denote these as s^μ, call them "timelike GPS coordinates", and define them by

$$s^\alpha = W^\alpha_\mu s^\mu. \tag{2.160}$$

This is a simple algebraic relabeling of the names of the four GPS coordinates, such that $s^{\mu=0}$ is timelike and $s^{\mu=a}$ is spacelike. In these coordinates, the gauge condition (2.159) reads

$$W^\alpha_\mu W^\alpha_\nu g^{\mu\nu}(s) = 0. \tag{2.161}$$

This can be interpreted geometrically as follows. The (timelike) GPS coordinates are coordinates s^μ such that the four 1-forms fields

$$\omega^\alpha = W^\alpha_\mu ds^\mu \tag{2.162}$$

are null.

Let us now jump from Minkowski space to full GR. Consider GR coupled with four satellites of negligible mass that move geodesically and whose world lines emerge from a point O with directions and velocity as above. Locally around O the metric can be taken to be Minkowskian; therefore the details of the initial conditions of the satellites worldlines can be taken as above. The phase space of this system is the one of pure GR plus 10 parameters, giving the location of O and the Lorentz orientation of the initial tetrahedron of velocities. The integration of the satellites' geodesics and of the light cones can be arbitrarily complicated in an arbitrary metric. However, if the metric is sufficiently regular, there will still be a region $\mathcal{R}$ in which the radio signals broadcasted by the satellites are received. (In case of multiple reception, the strongest one can be selected. That is, if the past light cone of p intersects l more than once, generically there will be one intersection which is at shorter luminosity distance.) Thus, we still have well defined physical coordinates s^α on $\mathcal{R}$. Equation (2.159) holds in these coordinates, because it depends just on the properties of the light propagation around p. We define also timelike GPS coordinates s^μ by (2.160), and we get the condition (2.161) on the metric tensor.

To study the evolution of the metric tensor in GPS coordinates it is easier to shift to ADM variables N, N^a, γ_{ab}. These are functions of the covariant components of the metric tensor, defined in general by

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu = N^2dt^2 - \gamma_{ab}(dx^a - N^a dt)(dx^b - N^b dt). \tag{2.163}$$

Equivalently, they are related to the contravariant components of the metric tensor by

$$g^{\mu\nu}v_\mu v_\nu = -\gamma^{ab}v_a v_b + (n^\mu v_\mu)^2, \tag{2.164}$$

where γ^{ab} is the inverse of γ_{ab} and $n^\mu = (1/N, N^a/N)$. Using these variables, the gauge condition (2.161) reads

$$W^\alpha_a W^\alpha_b \gamma^{ab} = (W^\alpha_\mu n^\mu)^2. \tag{2.165}$$

Notice now that this can be solved for the Lapse and Shift as a function of the three-metric (recall that W^α_μ are fixed numbers), obtaining

$$n^\mu = W^\mu_\alpha q^\alpha \tag{2.166}$$

where W^μ_α is the inverse of the matrix W^α_μ and

$$q^\alpha = \sqrt{W^\alpha_a W^\alpha_b \gamma^{ab}}. \tag{2.167}$$

Or, explicitly,

$$N = \frac{1}{W^0_\alpha q^\alpha}, \qquad N^a = \frac{W^a_\alpha q^\alpha}{W^0_\alpha q^\alpha}. \tag{2.168}$$

The geometrical interpretation is as follows. We want the 1-form ω^α defined in (2.162) to be null, namely its norm to vanish. But in the ADM formalism this norm is the sum of two parts: the norm of the pull back of ω^α on the constant time ADM surface, which is q^α, given in (2.167), and depends on the three metric; plus the square of the projection of ω^α on n^μ. We can thus obtain the vanishing of the norm by adjusting the Lapse and Shift. We have four conditions (one per each α) and we can thus determine Lapse and Shift out of three metric. In other words, whatever is the three metric, we can always adjust Lapse and Shift so that the gauge condition (2.161) is satisfied. But in the ADM formalism, the arbitrariness of the evolution in the Einstein equations is entirely captured by the freedom in choosing Lapse and Shift. Since here Lapse and Shift are uniquely determined by the three metric, evolution is determined uniquely if the initial data on a Cauchy surface are known. Therefore the evolution in the GPS coordinate s^0 of the GPS components of the metric tensor, $g_{\mu\nu}(s)$, is governed by deterministic equations: the ADM evolution equation with Lapse and Shift determined by equations (2.167–2.168). Notice also that evolution is local, since the ADM evolution equations, as well as the equations (2.167–2.168), are local.[27]

How can the evolution of the quantities $g_{\mu\nu}(s)$ be local? The conditions on the null surfaces described in the previous paragraph are nonlocal. Coordinate distances yield typically to nonlocality: Imagine we define physical coordinates in the solar system using the cosmological time t_c and the spatial distances x_S, x_E, x_J (at fixed t_c) from, say, the Sun the Earth and Jupiter. The metric tensor $g_{\mu\nu}(t_c, x_S, x_E, x_J)$ in these coordinates is a gauge invariant observable, but its evolution is highly non local. To see this, imagine that in this moment (in cosmological time), Jupiter is swept away by a huge comet. Then the value of $g_{\mu\nu}(t_c, x_S, x_E, x_J)$ here changes instantaneously, without any local cause: the value of the coordinate x_J has changed because of an event happened far away. What's special about the GPS coordinates that avoids this nonlocality? The answer is that the value of a GPS coordinate in a point p does in fact depend on what happens "far away" as well. Indeed, it depends on what happens to the satellite. However, it only depends on what happened to the satellite when it was broadcasting the signal received in p, and this is in the past of p ! If p is in the past domain of dependence of a partial Cauchy surface Σ, then the value of $g_{\mu\nu}(s)$ in p is completely determined by the metric and its derivative on Σ, namely evolution is causal, because the entire information needed to set up the GPS coordinates is in the data in Σ. See Figure 2.5. Explicitly, the $s^\alpha = constant$ surfaces around Σ can be uniquely integrated ahead all the way to p. They certainly can, as they represent just the evolution of a light front! This is how local evolution is achieved by these coordinates.

Summarizing, I have introduced a set of physical coordinates, determined by certain material bodies. Geometrical quantities such as the components of the metric tensor expressed in physical coordinates are of gauge invariant observables. There is no need to introduce a large unrealistic amount of matter or to construct complicated and unrealistic physical quantities out of the metric tensor. Four particles are sufficient to coordinatize a (region of a) four-geometry. Furthermore, the coordinatization procedure is not artificial: it is the real one utilized by existing technology.

The components of the metric tensor in (timelike) GPS coordinates can be measured as follows. Take a rod of physical length L (small with respect to the distance along which the gravitational field changes significantly) with two GPS devices at its ends (reading timelike GPS coordinates). Orient the rod (or search among recorded readings) so that the two GPS devices have the same reading s of all coordinates except for s^1. Let δs^1 be the difference in the two s^1 readings. Then we have along the rod

$$ds^2 = g_{11}(s)\delta s^1 \delta s^1 = L^2. \tag{2.169}$$

Therefore

$$g_{11}(s) = \left(\frac{L}{\delta s^1}\right)^2. \tag{2.170}$$

Non-diagonal components of $g_{ab}(s)$ can be measured by simple generalizations of this procedure. The $g_{0b}(s)$ are then algebraically determined by the gauge conditions. In a thought experiment, a spaceship could travel in a spacetime region and compose a map of values of the GPS components of the metric tensor. Instead of using a rod, which is a rather crude device for measuring distances, one could send a light pulse forward and back between the two GPS devices kept at fixed spatial s^μ coordinates. If T is the (physical) time for flying back and forward measured by a precise clock on one device, then $g_{11}(s) = \left(cT/2\delta s^1\right)^2$. This is valid as far as T and L are small compared to the distances over which $g_{ab}(s)$ changes by an amounts of the order of the experimental errors.

The individual components of the metric tensor expressed in physical coordinates are measurable. The statement that "the curvature is measurable but the metric is not measurable", which is often heard, is nonsense. Both metric and curvature in physical coordinates are measurable and predictable. Neither metric nor curvature in arbitrary nonphysical coordinates, are measurable.

---

[27] This does not imply that the full set of equations satisfied by $g_{\mu\nu}(s)$ must be local, since initial conditions on $s^0 = 0$ satisfy four other constraints besides the ADM ones.

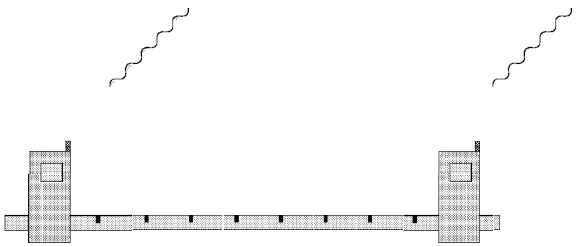


Figure 2.6: A simple apparatus to measure the gravitational field. Two GPS devices, reading s^μ_L and s^μ_R respectively, connected by a 1 meter long rod. If for instance $s^\mu_R = s^\mu_L$ for $\mu = 0, 2, 3$, then the local value of $g_{11}(s)$ is: $g_{11}(s) = (s^\mu_R - s^\mu_L)^{-2} m^2$.

The GPS coordinates are partial observables (see chapter 3). The complete observables are the quantities $g_{\mu\nu}(s)$, for any given value of the coordinates s^μ. These quantities are diffeomorphism invariant, are uniquely determined by the initial data and in a canonical formulation are represented by functions on the phase space that commute with all constraints.

The GPS observables are a straightforward generalization of Einstein's "spacetime coincidences". In a sense, they are *precisely* Einstein's point coincidences. Einstein's "material points" are just replaced by photons (light pulses): the spacetime point s^α is characterized as the meeting point of four photons designated by the fact of carrying the radio signals s^α.

Bibliographical note

There are many beautiful classic textbook on GR. Two among the best, offering remarkably different points of view on the theory, are Weinberg [67] and Wald [68]. The first stresses the similarity between GR and flat space field theory; the second, on the contrary, emphasizes the geometrical reading of GR. Here I have followed a third path: emphasis on the change of the notions of space and time needed for general relativistic physics (which affects quantization dramatically), but little emphasis on the geometrical interpretation of the gravitational field (which is going to be largely lost in the quantum theory).

Relevant mathematics is nicely presented, for instance, in the text by Choquet-Bruhat, DeWitt-Morette and Dillard-Bleick [69] and in [14]. On the large empirical evidence in favor of GR, piled up in the recent years, see Ciufolini and Wheeler [70].

The tetrad formalism is mainly due to Cartan and Weyl. The first order formalism to Palatini. The Plebanski two-form was introduced in [71]. The selfdual connection, which is a the roots of Ashtekar's canonical theory (see chapter 4), was introduced by Amitaba Sen [72]. The lagrangian formulation for the selfdual connections was given in [73]. A formulation of GR based on the sole connection is discussed in [74].

Interesting reconstructions of Einstein path towards GR are in [75, 76]. Kretschmann's objection to the significance of general covariance appeared in [77]. On this, see also Anderson's book [78]. An account of the historical debate on the interpretation of space and motion is in Julian Barbour's [79], a wonderful historical book. In philosophy of science, the debate was reopened by a 1987 paper

on the hole argument by John Earman and John Norton [80]. On the contemporary version of this debate, see [81, 82, 83, 84]. On the physical side of the discussion of what is "observable" in GR, see [85].

The discussion of the different meaning of time follows [86]. A surprising and inspiring book on the subject is Fraser [87], a book that will convince the reader that the notion of time is far from being a monolithic concept. The literature on the problem of time in quantum gravity is vast. I list only a few pointers here, distinguishing various problems. Origin of the "arrow of time" and the cosmological time asymmetry [88]; disappearance of the coordinate-time variable in canonical quantum gravity [89]; possibility of a consistent interpretation of quantum mechanics for systems without global time [24, 90, 91]; problems in choosing an "internal time" in general relativity, and the properties that such an internal time should have [92]; see also [93]. The presentation of the GPS observables follows [94]; see also [95, 96].

Chapter 3

Relativistic mechanics

In its conventional formulation, mechanics describes the evolution of states and observables in time. This evolution is governed by a Hamiltonian. This is also true for special relativistic theories: evolution is governed by a representation of the Poincaré group, which includes a Hamiltonian. This conventional formulation is not sufficiently broad, because general relativistic systems –in fact, the world in which we live– do not fit in this conceptual scheme. Therefore we need a more general formulation of mechanics than the conventional one. This formulation must be based on notions of "observable" and "state" that maintain a clear meaning in a general relativistic context. A formulation of this kind is described in this chapter.

The conventional structure of conventional nonrelativistic mechanics already points out rather directly to the relativistic formulation described here. Indeed, many aspects of this formulation are already utilized by many authors. For instance, Arnold [97] identifies the (presymplectic) space with coordinates (t, q^i, p_i) (time, lagrangian variables and their momenta) as the natural home for mechanics. Souriau has developed a beautiful and little known relativistic formalism [98]. Probably the first to consider the point of view used here was Lagrange himself, in pointing out that the most convenient definition of "phase space" is the space of the physical motions [99]. Many of the tools used below are also used in hamiltonian treatments of general covariant theories as constrained systems, although generally within a rather obscure interpretative cloud.

3.1 Non-relativistic mechanics: *Mechanics is about time evolution*

I begin with a brief review of conventional mechanics. This is useful to fix notations and to introduce some notions that will play a role in the relativistic formalism. I give no derivations here: they are standard, and they can be obtained as a special case of the derivations in the next section.

Lagrangian. A dynamical system with m degrees of freedom describes the evolution in time t of m lagrangian variables q^i, where $i = 1, \dots, m$. The space in which the variables q^i take value is the m-dimensional (nonrelativistic) configuration space $\mathcal{C}_0$. The dynamics of the system is determined by a single function of $2m$ variables $L(q^i, v^i)$, the Lagrangian. Given two times t_1 and t_2 and two points q_1^i and q_2^i in $\mathcal{C}_0$, physical motions are such that the action

$$S[q] = \int_{t_1}^{t_2} dt\; L\left(q^i(t), \frac{dq^i(t)}{dt}\right) \tag{3.1}$$

is an extremum in the space of the motions $q^i(t)$ such that $q^i(t_1) = q_1^i$ and $q^i(t_2) = q_2^i$. A dynamical system is therefore specified by the couple $(\mathcal{C}_0, L)$. Physical motions satisfy the Lagrange equations

$$\frac{d}{dt}\; p_i\left(q^i(t), \frac{dq^i(t)}{dt}\right) = F_i\left(q^i(t), \frac{dq^i(t)}{dt}\right) \tag{3.2}$$

where momenta and forces are defined by

$$p_i(q^i, v^i) = \frac{\partial L(q^i, v^i)}{\partial v^i}, \qquad F_i(q^i, v^i) = \frac{\partial L(q^i, v^i)}{\partial q^i}. \tag{3.3}$$

Hamiltonian. The Lagrange equations can be cast in first order form by using the lagrangian coordinates q^i and the momenta p_i as variables. Inverting the function $p_i(q^i, v^i)$ yields the function $v^i(q^i, p_i)$; inserting this in the function $F_i(q^i, v^i)$ defines the force $f_i(q^i, p_i) \equiv F_i(q^i, v^i(q^i, p_i))$ as functions of coordinate and momenta. The equations of motion (3.2) become

$$\frac{dq^i(t)}{dt} = v^i(q^i(t), p_i(t)), \qquad \frac{dp_i(t)}{dt} = f_i(q^i(t), p_i(t)). \tag{3.4}$$

These equations are determined by the function $H_0(q^i, p_i)$, the nonrelativistic Hamiltonian defined as $H_0(q^i, p_i) = p_i v^i(q^i, p_i) - L(q^i, v^i(q^i, p_i))$. Indeed, (3.4) is equivalent to (3.2) with

$$v^i(q^i, p_i) = \frac{\partial H_0(q^i, p_i)}{\partial p_i}, \qquad f_i(q^i, p_i) = -\frac{\partial H_0(q^i, p_i)}{\partial q^i}. \tag{3.5}$$

Symplectic. The Hamilton equations (3.4–3.5) can be written in a useful and compact geometric language. The $2m$-dimensional space coordinatized by the coordinates q^i and the momenta p_i is the nonrelativistic phase space Γ_0 (The reason for the subscript $_0$ will be clear below.) Time evolution is a flow $(q^i(t), p_i(t))$ in this space; the vector field on Γ_0 tangent to this flow is

$$X_0 = v_i(q^i, p_i)\frac{\partial}{\partial q^i} + f_i(q^i, p_i)\frac{\partial}{\partial p_i}. \tag{3.6}$$

Therefore the dynamics is specified by assigning the vector field X_0 on Γ_0. Now, Γ_0 can be interpreted as the cotangent space $T^*\mathcal{C}_0$. Any cotangent space carries a natural one-form $\theta_0 = p_i dq^i$.[1] A space equipped with a preferred one-form has the remarkable property that every function f determines a vector field X_f via the relation $(d\theta_0)(X_f) = -df$. A straightforward calculation shows that the flow defined by H_0 is precisely the time evolution vector field (3.6). Therefore the equations of motion (3.4-3.5) can be written simply as

$$(d\theta_0)(X_0) = -dH_0. \tag{3.7}$$

The two-form $\omega_0 = d\theta_0$ entering (3.7) is symplectic [2]. A dynamical system is determined by a triple $(\Gamma_0, \omega_0, H_0)$, where Γ_0 is a manifold, ω_0 is a symplectic two-form and H_0 is a function on Γ_0.

Presymplectic. A very elegant formulation of mechanics, and a crucial step in the direction of the relativistic theory, is provided by the presymplectic formalism. This formalism is based on the idea of describing motions by using the graph of the function $(q^i(t), p_i(t))$ instead than the functions themselves. The graph of the function $(q^i(t), p_i(t))$ is a unparametrized curve $\tilde{\gamma}$ in the $(2m+1)$-dimensional space $\Sigma = R \times \Gamma_0$, with coordinates (t, q^i, p_i); it is formed by all the points $(t, q^i(t), p_i(t))$ in this space. The vector field

$$X = \frac{\partial}{\partial t} + v(q^i, p_i)\frac{\partial}{\partial q^i} + f(q^i, p_i)\frac{\partial}{\partial p^i} \tag{3.8}$$

[1] It is defined intrinsically by $\theta_0(X)(s) = s(\pi X)$ where X is a vector field on $T^*\mathcal{C}_0$, s a point in $T^*\mathcal{C}_0$ and π the bundle projection.

[2] That is, closed and nondegenerate. Closed means $d\omega_0 = 0$, nondegenerate means that $\omega_0(X) = 0$ implies $X = 0$.

is tangent to all these curves. (So is any other vector field obtained scaling X, namely any vector field $X' = fX$, where f is a scalar function on Σ.) Now, consider the Poincaré one-form

$$\theta = p_i dq^i - H_0(q^i, p_i)dt. \tag{3.9}$$

on Σ. The two-form $\omega = d\theta$ is closed but it is not nondegenerate (no two-form is nondegenerate in odd dimensions); that is, there is a vector field X (called the null vector field of ω) satisfying

$$(d\theta)(X) = 0. \tag{3.10}$$

The integral curves[3] of the null vector field of a two-form ω are called the "orbits" of ω. It is easy to see that X given in (3.8) satisfies (3.10). Therefore the graphs of the motions are simply the orbits of $d\theta$. In other words, (3.10) is a rewriting of the equations of motion.

A space Σ equipped with a closed degenerate two-form ω is called presymplectic. A dynamical system is thus completely defined by a presymplectic space (Σ, ω). We use also the notation (Σ, θ), where $\omega = d\theta$.

Notice that equation (3.10) is homogeneous and therefore it determines X only up to scaling. This is consistent with the fact that the vector field tangent to the motions is defined only up to scaling. That is, with the fact that motions are represented by *unparametrized* curves in Σ.

Finally, it is easy to see that the action (3.1) is simply the line integral of the Poincaré one-form (3.9) along the orbits: if $\tilde{\gamma}$ is an orbit $(t, q^i(t), p_i(t))$ of ω, then the action of the motion $q^i(t)$ is

$$S[q] = \int_{\tilde{\gamma}} \theta. \tag{3.11}$$

Extended. Finally, let me come to a formulation of dynamics that extends naturally to general relativistic systems. In the light of the presymplectic formulation described above, it is natural to consider the relativistic configuration space

$$\mathcal{C} = R \times \mathcal{C}_0 \tag{3.12}$$

coordinatized by the $m+1$ variables (t, q^i) and to describe motions with the graphs of the functions $q^i(t)$, which are unparametrized curves in $\mathcal{C}$. Consider the cotangent space $T^*\mathcal{C}$, with coordinates (t, q^i, p_t, p_i) and the function

$$H(t, q^i, p_t, p_i) = p_t + H_0(q^i, p_i) \tag{3.13}$$

on this space. Let Σ be the surface in $T^*\mathcal{C}$ defined by

$$H(q^i, t, p_i, p_t) = 0. \tag{3.14}$$

We can coordinatize Σ with the coordinates (t, q^i, p_i). Since it is a cotangent space, $T^*\mathcal{C}$ carries a natural one-form, which is

$$\tilde{\theta} = p_i dq^i + p_t dt. \tag{3.15}$$

The restriction of this one-form to the surface (3.14) is precisely (3.9). Therefore the surface (3.14) is the presymplectic space that defines the dynamics.

In other words, the dynamics is completely defined by the couple $(\mathcal{C}, H)$: a relativistic configuration space $\mathcal{C}$ and a function H on $T^*\mathcal{C}$. The graphs of the motions are simply the orbits of $d\tilde{\theta}$ on the surface (3.14).[4] I call H the relativistic hamiltonian.

---

[3] An integral curve of a vector field is a curve everywhere tangent to the field.

[4] More precisely, the projections of these orbits on $\mathcal{C}$.

Remarkably, the dynamics can be directly expressed in terms of a variational principle based on $(\mathcal{C}, H)$: An unparametrized curve γ in $\mathcal{C}$ describes a physical motion if $\tilde{\gamma}$ extremizes the integral

$$S[\tilde{\gamma}] = \int_{\tilde{\gamma}} \tilde{\theta} \tag{3.16}$$

in the class of the curves $\tilde{\gamma}$ in $T^*\mathcal{C}$ satisfying (3.14) whose restriction γ to $\mathcal{C}$ connects two given points (t_1, q_1^i) and (t_2, q_2^i).

The relativistic configuration space $\mathcal{C}$ has the structure (3.12) and the relativistic hamiltonian H has the form (3.13). As we shall see, the structure (3.12–3.13) does not survive in the relativistic formulation of mechanics.

Relativistic phase space. Denote Γ the space of the orbits of $d\theta$ in Σ. There is a natural projection $\pi : \Sigma \to \Gamma$ that sends each point of Σ to the curve to which it belong. It is not hard to show that there is one and only one symplectic two-form ω_{ph} on Γ such that its pull back to Σ is $d\theta$, namely $\pi^*\omega_{ph} = d\theta$. Therefore Γ is a symplectic space. Γ is the space of the physical motions; I shall call it the relativistic phase space.

The relation between the relativistic phase space Γ and the nonrelativistic phase space $\Gamma_0 = T^*\mathcal{C}_0$ is the following. Γ_0 is the space of the instantaneous states: the states that the system can have at a fixed time $t = t_0$. On the other hand, Γ is the space of all solutions of the equations of motions. Now, fix a time, say $t = t_0$. If at $t = t_0$ the system is in an initial state in Γ_0 it will then evolve in a well defined motion. The other way around, each motion determines an instantaneous state at $t = t_0$. Therefore there is a one-to-one mapping between Γ and Γ_0. The identification between Γ and Γ_0 depends on the t_0 chosen.

Hamilton-Jacobi. The Hamilton-Jacobi equation is

$$\frac{\partial S(q^i, t)}{\partial t} + H_0\left(q^i, \frac{\partial S(q^i, t)}{\partial q^i}\right) = 0. \tag{3.17}$$

If a family of solution $S(q^i, Q^i, t)$ depending in m parameters Q_i is found, then we can compute the function

$$P_i(q^i, Q^i, t) \quad = \quad -\frac{\partial S(q^i, Q^i, t)}{\partial Q^i}. \tag{3.18}$$

by simple derivation. Inverting this function we obtain

$$q^i(t) = q^i(Q^i, P_i, t) \tag{3.19}$$

which are physical motions, namely the general solution of the equation of motion, where the quantities (Q^i, P_i) are the $2m$ integration constants.

Solutions of (3.17) can be found in the form $S(q^i, Q^i, t) = Et - W(q^i, Q^i)$ where W satisfies

$$H_0\left(q^i, \frac{\partial W(q^i, Q^i)}{\partial q^i}\right) = E. \tag{3.20}$$

S is called the principal Hamilton-Jacobi function, W is called the characteristic Hamilton-Jacobi function,

The Hamilton-Jacobi equation (3.17) can be obtained from the classical limit of the Schrödinger equation.

The Hamilton function. Consider two points (t_1, q_1^i) and (t_2, q_2^i) in $\mathcal{C}$. The function on $\mathcal{G} = \mathcal{C} \times \mathcal{C}$

$$S(t_1, q_1^i, t_2, q_2^i) = \int_{t_1}^{t_2} dt\ L(q^i(t), \dot{q}^i(t)) \tag{3.21}$$

where $q^i(t)$ is the physical motion from $q_1^i(t_1)$ to $q_2^i(t_2)$ (that minimizes the action), is called the Hamilton function. Equivalently,

$$S(t_1, q_1^i, t_2, q_2^i) = \int_{\tilde{\gamma}} \theta, \tag{3.22}$$

where $\tilde{\gamma}$ the orbit in Σ that projects to $q^i(t)$. Notice the difference between the action (3.1) and the Hamilton function (3.21): the first is a functional of the motion; the second is a function of the end points. It is not hard to see that the Hamilton function solves the Hamilton-Jacobi equation (in both sets of variables). The Hamilton function is therefore a preferred solution of the Hamilton-Jacobi equation. If we know the Hamilton function, we have solved the equations of motion, because we obtain the general solutions of the equation of motion in the form $q^i = q^i(t, Q^i, P_i, T)$ by simply inverting the function

$$P_i(t, q^i, T, Q^i) = \frac{\partial S(t, q^i, T, Q^i)}{\partial Q^i} \tag{3.23}$$

with respect to q^i. The resulting function $q^i(t, T, Q^i, P_i)$ is the general solution of the equations of motion where the integration constants are the initial coordinate and momenta Q^i, P_i at time T.

Thus, the action defines a dynamical system; the Hamilton function directly gives all the motions.[5] The Hamilton function (3.21) is the classical limit of the quantum mechanical propagator.

Example: a pendulum. Let α be the lagrangian variable describing the elongation of a simple harmonic oscillator, which I call "pendulum" for simplicity. The lagrangian is $L(\alpha, v) = \frac{mv^2}{2} - \frac{m\omega^2\alpha^2}{2}$; the nonrelativistic Hamiltonian is $H_0(\alpha, p) = \frac{p^2}{2m} + \frac{m\omega^2\alpha^2}{2}$. The extended configuration space has coordinates (t, α) and the relativistic hamiltonian is

$$H(t, \alpha, p_t, p) = p_t + \frac{p^2}{2m} + \frac{m\omega^2\alpha^2}{2}. \tag{3.24}$$

Choose coordinates (t, α, p) on the constraint surface $H = 0$, which is therefore defined by $p_t = -H_0(\alpha, p)$. The restriction of the one-form $\tilde{\theta} = p_t\, dt + p\, d\alpha$ to this surface is

$$\theta = p\, d\alpha - \left(\frac{p^2}{2m} + \frac{m\omega^2\alpha^2}{2}\right) dt. \tag{3.25}$$

The presymplectic two-form is therefore

$$\omega = d\theta = dp \wedge d\alpha - \frac{p}{m} dp \wedge dt - m\omega^2\alpha\ d\alpha \wedge dt. \tag{3.26}$$

The orbits are obtained by integrating the vector field

$$X = X_t \frac{\partial}{\partial t} + X_\alpha \frac{\partial}{\partial \alpha} + X_p \frac{\partial}{\partial p} \tag{3.27}$$

satisfying $\omega(X) = 0$. Inserting (3.26) and (3.27) in $\omega(X) = 0$ we get

$$\begin{aligned} \omega(X) &= X_t \left(-\frac{p}{m} dp - m\omega^2\alpha\, d\alpha\right) + X_\alpha \left(dp + m\omega^2\alpha\, dt\right) + X_p \left(-d\alpha + \frac{p}{m}\, dt\right) \\ &= \left(-\frac{p}{m} X_t + X_\alpha\right) dp + \left(-m\omega^2\alpha X_t - X_p\right) d\alpha + \left(m\omega^2\alpha X_\alpha + \frac{p}{m} X_p\right) dt \\ &= 0. \end{aligned} \tag{3.28}$$

[5] Hamilton (talking about himself in the third person): "Mr. Lagrange's function *states* the problem, Mr. Hamilton's function *solves* it" [100].

Writing $\frac{dt(\tau)}{d\tau} = X_t$, $\frac{d\alpha(\tau)}{d\tau} = X_\alpha$, $\frac{dp(\tau)}{d\tau} = X_p$, equation (3.28) reads

$$\frac{d\alpha(\tau)}{d\tau} - \frac{p}{m}\frac{dt(\tau)}{d\tau} = 0, \qquad -\frac{dp(\tau)}{d\tau} - m\omega^2\alpha\frac{dt(\tau)}{d\tau} = 0, \tag{3.29}$$

and a third equation dependent from the first two. (3.29) can be written as

$$\frac{d\alpha(t)}{dt} = \frac{p}{m}, \qquad \frac{dp(t)}{dt} = -m\omega^2\alpha, \tag{3.30}$$

which are the Hamilton equations of the pendulum. We can write its general solution in the form

$$\alpha(t) = a\, e^{i\omega t} + \bar{a}\, e^{-i\omega t}. \tag{3.31}$$

The Hamilton function $S(\alpha_1, t_1, \alpha_2, t_2)$ is the preferred solution of the Hamilton-Jacobi equation

$$\frac{\partial S(\alpha,t)}{\partial t} + \frac{1}{2m}\left(\frac{\partial S(\alpha,t)}{\partial \alpha}\right)^2 + \frac{m\omega^2\alpha^2}{2} = 0, \tag{3.32}$$

obtained computing the action of the physical motion $\alpha(t)$ that goes from $\alpha(t_1) = \alpha_1$ to $\alpha(t_2) = \alpha_2$. This motion is given by (3.31) with

$$a = \frac{\alpha_1 e^{-i\omega t_2} + \alpha_2 e^{-i\omega t_1}}{2i\sin(\omega(t_1 - t_2))}. \tag{3.33}$$

Inserting this in the action and integrating we obtain the Hamilton function

$$S(\alpha_1, t_1, \alpha_2, t_2) = m\omega\, \frac{2\alpha_1\alpha_2 - (\alpha_1^2 + \alpha_2^2)\cos(\omega(t_1 - t_2))}{2\sin(\omega(t_1 - t_2))}. \tag{3.34}$$

This concludes the short review of nonrelativistic mechanics. I now consider the generalization of this formalism to relativistic systems.

3.2 Relativistic mechanics

3.2.1 Structure of relativistic systems: partial observables, relativistic states

Is there a version of the notions of "state" and "observable" broad enough to apply naturally to relativistic systems? I begin by introducing the main notions and tools of covariant mechanics in the context of a simple system.

The pendulum revisited. Say we want to describe the small oscillations of a pendulum. To this aim, we need *two* measuring devices. A clock and a device that reads the elongation of the pendulum. Let t be the reading of the clock (in seconds) and α the reading of the device measuring the elongation of the pendulum (in centimeters). Call the variables t and α the *partial observables* of the pendulum. (I use also *relativistic observables*, or simply *observables*, if there is no risk of confusion with the nonrelativistic notion of observable, which is different.)

A useful observation is a reading of the time t and the elongation α, *together*. Thus, an observation yields a pair (t, α). Call a pair obtained in this manner an *event*.

Let $\mathcal{C}$ be the two-dimensional space with coordinates t and α. Call $\mathcal{C}$ the *event space* of the pendulum. (I use also *relativistic configuration space*, or simply *configuration space*, if there is no risk of confusion with the nonrelativistic configuration space $\mathcal{C}_0$, which is different.)

Experience shows we can find mathematical laws characterizing *sequences* of events. This is the reason we can do science. These laws have the following form. Call a unparametrized curve γ in $\mathcal{C}$ a *motion* of the system. Perform a sequence of measurements of pairs (t, α), and find that the points representing the measured pairs sit on a motion γ. Then we say that γ is a *physical motion*. We express a motion as a relation in $\mathcal{C}$

$$f(\alpha, t) = 0. \tag{3.35}$$

Thus a motion γ is a relation, or a *correlation*, between partial observables.

Then, disturb the pendulum (push it with a finger) and repeat the entire experiment over. At each repetition of the experiment, a different motion γ is found. That is, a different mathematical relation of the form (3.35) is found. Experience shows that the space of the physical motions is very limited: it is just a two-dimensional space. Only a two-dimensional space of curves γ is realized in nature.

In the case of the small oscillations of a frictionless pendulum, we can coordinatize the physical motions by the two real numbers $A \geq 0$ and $0 \leq \phi < 2\pi$, and (3.35) is given by

$$f(\alpha, t; A, \phi) = \alpha - A\sin(\omega t + \phi) = 0. \tag{3.36}$$

This equation gives a curve γ in $\mathcal{C}$ for each couple (A, ϕ).

Let Γ be the two-dimensional space of the physical motions, coordinatized by A and ϕ. Γ is the *relativistic phase space* of the pendulum (or the *space of the motions*). A point in Γ, is also called a *relativistic state*. (Or a *Heisenberg state*, or simply a *state*, if there is no risk of confusion with the nonrelativistic notion of state, which is different.)

Equation (3.36) is the mathematical law that captures the empirical information we have on the pendulum. This equation is the *evolution equation* of the system. The function f is the *evolution function* of the system.

A relativistic state is determined by a couple (A, ϕ). It determines a curve γ in the (t, α) plane. That is, it determines a correlations between the two partial observables t and α, via equation (3.36). If we disturb the pendulum by interacting with it, or if we start a new experiment over, we have a new state. The state remains the same if we observe the pendulum and the clock without disturbing them (here we disregard quantum theory, of course).

Summarizing: *each state in the phase space Γ determines a correlation between the observables in the configuration space $\mathcal{C}$.* The set of these relations is captured by the evolution equation (3.36), namely by the vanishing of a function

$$f : \Gamma \times \mathcal{C} \to R. \tag{3.37}$$

The evolution equation $f = 0$ expresses all predictions that can be made using the theory. Equivalently, these predictions are captured by the surface (3.37) in the Cartesian product of the phase space with the configuration space.

General structure of the dynamical systems. The $(\mathcal{C}, \Gamma, f)$ language described above is general. On the one hand, it is sufficient to describe all predictions of conventional mechanics. On the other hand, it is broad enough to describe general relativistic systems. All fundamental systems can be described (to the accuracy at which quantum effects can be disregarded) by making use of these concepts:

(i) The relativistic *configuration space* $\mathcal{C}$, of the partial *observables.*

(ii) The relativistic *phase space* Γ of the relativistic *states.*

(iii) The *evolution equation* $f = 0$, where $f : \Gamma \times \mathcal{C} \to V$.

V is a linear space. The state in the phase space Γ is fixed until the system is disturbed. Each state in Γ determines (via $f = 0$) a motion γ of the system, namely a relation, or a set of relations, between the observables in $\mathcal{C}$.

A motion is not necessarily a one-dimensional curve in $\mathcal{C}$: it can be a surface in $\mathcal{C}$ of any dimension k. If $k > 1$, we say that there is gauge invariance. For a system with gauge invariance

we call "motion", equivalently, the motion itself and any curve within it. In this chapter we shall not deal much with systems with gauge invariance, but we shall mention them where relevant.

Predictions are obtained as follows. We first perform enough measurements to find out the state. (In reality the state of a large system is often "guessed" on the basis of incomplete observations and reasonable assumptions, justified inductively.) Once the state is so determined or guessed, the evolution equation predicts all the possible events, namely all the allowed correlations between the observables, in any subsequent measurement.

In the example of the pendulum, for instance, the equation predicts the value of α that can be measured together with any given t, or the values of t that can be measured together with any given α. These predictions are valid until the system is disturbed.

The definitions of observable, state, configuration space and phase space given here are different from the conventional definition. In particular, notions of instantaneous state, evolution in time, observable at a fixed time, play no role here. These notions make no sense in a general relativistic context. For nonrelativistic systems, the usual notions can be recovered from the definitions given. The relation between the relativistic definitions considered here and the conventional nonrelativistic notions is discussed below in section 3.2.4.

The task of mechanics is to find the $(\mathcal{C}, \Gamma, f)$ description for all physical systems. The first step, kinematics, consists in the specification of the observables that characterize the system. Namely the specification of the configuration space $\mathcal{C}$ and its physical interpretation. Physical interpretation means the association of coordinates on $\mathcal{C}$ with measuring devices. The second step, dynamics, consists in finding the phase space Γ and the function f that describe the physical motions of the system.

In the next section, I describe a relativistic hamiltonian formalism for mechanics, based on the relativistic notions of state and observable defined here.

3.2.2 Hamiltonian mechanics

Elementary physical systems can be described by hamiltonian mechanics.[6] Once the kinematics –that is, the space $\mathcal{C}$ of the partial observables q^a– is known, the dynamics –that is, Γ and f– is fully determined by giving a surface Σ in the space Ω of the observables q^a and their momenta p_a. The surface Σ can be specified by giving a function $H : \Omega \to R^k$. Σ is then defined by $H = 0$.[7] Denote $\tilde{\gamma}$ a curve in Ω (observables and momenta) and γ its restriction to $\mathcal{C}$ (observables alone). H determines the physical motions via the following

> ***Variational principle.*** *A curve γ connecting the events q_1^a and q_2^a is a physical motion if $\tilde{\gamma}$ extremizes the action*
>
> $$S[\tilde{\gamma}] = \int_{\tilde{\gamma}} p_a \, dq^a \tag{3.38}$$
>
> *in the class of the curves $\tilde{\gamma}$ satisfying*
>
> $$H(q^a, p_a) = 0 \tag{3.39}$$
>
> *whose restriction γ to $\mathcal{C}$ connects q_1^a and q_2^a.*

All (relativistic and nonrelativistic) hamiltonian systems can be formulated in this manner.

[6] Perhaps because they are the classical limit of a quantum system.

[7] Different H's that vanish on the same surface Σ define the same physical system.

If $k = 1$, H is a scalar function, and is sometimes called the hamiltonian constraint. The case $k > 1$ is the case in which there is gauge invariance. In this case, the system (3.39) is sometimes called the system of the "constraint equations". I call H the *relativistic hamiltonian*, or, if there is no ambiguity, simply the *hamiltonian*. I denote the pair $(\mathcal{C}, H)$ as a *relativistic dynamical system*. The generalization to field theory is discussed in section 3.3.

The relativistic hamiltonian H is related to, but should not be confused with, the usual non-relativistic hamiltonian, denoted H_0 in this book. H always exists, while H_0 exists only for the nonrelativistic systems.

Indeed, notice that this formulation of mechanics is similar to the extended formulation of non-relativistic mechanics defined in section (3.1). The novelty is that $\mathcal{C}$ and H do not have the structure (3.12–3.13). The discussion above shows that this structure is *not* necessary in order to have a well-defined physical interpretation of the formalism. A nonrelativistic system is characterized by the fact that one of its partial observables q^a is singled out as having the special role of independent variable t. This does not happen in a relativistic system. The following simple example shows that the relativistic formulation of mechanics is a proper generalization of standard mechanics.

Timeless double pendulum. I now introduce a genuinely timeless system, which I will repeatedly use as a simple model to illustrate the theory. Consider a mechanical model with two partial observables, say a and b, whose dynamics is defined by the relativistic hamiltonian

$$H(a, b, p_a, p_b) = \frac{1}{2}\ \left(p_a^2 + p_b^2 + a^2 + b^2 - 2E\right), \tag{3.40}$$

where E is a constant. The extended configuration space is $\mathcal{C} = R^2$. The constraint surface has dimension 3; it is the sphere of radius $\sqrt{2E}$ in $T^*\mathcal{C}$. The phase space has dimension 2. The motions are curves in the (a, b) space. For each state, the theory predicts the correlations between a and b.

A straightforward calculation (see below) shows that the evolution equation determined by H is the one of an ellipses in the (a, b) space

$$f(a, b; \alpha, \beta) = \left(\frac{a}{\sin\alpha}\right)^2 + \left(\frac{b}{\cos\alpha}\right)^2 + 2\frac{a}{\sin\alpha}\frac{b}{\cos\alpha}\cos\beta = 2E^2\sin^2\beta \tag{3.41}$$

where α and β parametrize Γ. Therefore motions are closed curves, in fact, ellipses, in $\mathcal{C}$. The system does not admit a conventional hamiltonian formulation, because for a nonrelativistic hamiltonian system motions in $\mathcal{C} = R \times \mathcal{C}_0$ are monotonic in $t \in R$ and therefore cannot be closed curves.

The example is not artificial. There exist cosmological models that have precisely this structure. For instance, we can identify a with the radius of a maximally symmetric universe and b with the spatially constant value of a field representing the matter content of the universe, and adopt the approximation in which these are the only two variables that govern the large scale evolution of the universe. Then the dynamics of general relativity reduces to a system with the structure (3.40).

The associated nonrelativistic system. The system (3.40) can also be viewed as follows. Consider a physical system, which we denote the "associated nonrelativistic system", which is a conventional nonrelativistic system, formed by two non-interacting harmonic oscillators. Let me stress that the associated non-relativistic system is a *different* physical system than the timeless double pendulum considered above. The timeless double pendulum has one degree of freedom, its associated non-relativistic system has two degrees of freedom. The partial observables of the associated non-relativistic system are the two elongations a and b, *and* the time t. The *nonrelativistic* hamiltonian that governs the evolution in t is

$$H_0(a, b, p_a, p_b) = \frac{1}{2}\ \left(p_a^2 + p_b^2 + a^2 + b^2 - 2E\right), \tag{3.42}$$

namely, it has the same form as the *relativistic* hamiltonian (3.40) of the the timeless double pendulum.[8] The constant term $2E$, of course, has no effect on the equations of motion; it only redefines the energy. Physically, we can view the relation between the two systems as follows. Imagine that we take the associated nonrelativistic system

[8]The *relativistic* hamiltonian of the associated nonrelativistic system is

$$H(a, b, t, p_a, p_b, p_t) = p_t + \frac{1}{2}\ \left(p_a^2 + p_b^2 + a^2 + b^2 - 2E\right). \tag{3.43}$$

but we decide to ignore the clock that measures t: we consider just measurements of the two observables a and b. Furthermore, assume that the energy of the double pendulum is constrained to vanish. Namely

$$\frac{1}{2}\left(p_a^2 + p_b^2 + a^2 + b^2\right) = E. \tag{3.44}$$

Then the observed relation between the measurements of a and b is described by the relativistic system (3.40).

Geometric formalism. As for nonrelativistic hamiltonian mechanics, the equations of motion can be expressed in an elegant geometric form. The variables (q^a, p_a) are coordinates on the cotangent space $\Omega = T^*\mathcal{C}$. Equation (3.39) defines a surface Σ in this space. The cotangent space carries the natural one-form

$$\tilde{\theta} = p_a dq^a. \tag{3.45}$$

Denote θ the restriction of $\tilde{\theta}$ to the surface Σ. The two-form $\omega = d\theta$ on Σ is degenerate: it has null directions. The integral surfaces of these null directions are the *orbits* of ω on Σ. Each such orbit projects from $T^*\mathcal{C}$ to $\mathcal{C}$ to give a surface in $\mathcal{C}$. These surfaces are the motions.

Consider the case $k = 1$. In this case Σ has dimension $2n - 1$, the kernel of ω is, generically, one-dimensional, and the motions are, generically, one-dimensional. Let $\tilde{\gamma}$, be a motion on Σ and X a vector tangent to the motion. Then

$$\omega(X) = 0. \tag{3.46}$$

To find the motions, we have just to integrate this equation. Equation (3.46) is the equation of motion. X is defined by the homogeneous equation (3.46) only up to a multiplicative factor. Therefore the tangent of the orbit is defined only up to a multiplicative factor, and therefore the parametrization of the the orbit is not determined by equation (3.46).

The case $k > 1$ is analogous. In this case Σ has dimension $2n - k$, the kernel of ω is, generically, k-dimensional and the motions are, generically, k-dimensional. X is then a k-dimensional multi-tangent, and it still satisfies equation (3.46).

Let $\pi : \Sigma \to \Gamma$ be the projection map that associate to each point of the constraint surface the motion to which the point belongs. The projection π equips the phase space Γ with a symplectic two-form ω_{ph}: this is defined as the two-form whose pull back to Σ under π is ω. Locally it exists and it is unique precisely because ω is degenerate along the orbits.

Relation with the variational principle. Let $\tilde{\gamma}$ be an orbit of ω on Σ, such that its restriction γ in $\mathcal{C}$ is bounded by the initial and final events q_1 and q_2. Let $\tilde{\gamma}'$ be a curve in Σ infinitesimally close to $\tilde{\gamma}$, and such that its restriction γ' is also bounded by q_1 and q_2. Let δs_1 (and δs_2) be the difference between the initial (and final) points of $\tilde{\gamma}$ and $\tilde{\gamma}'$. The four curves $\tilde{\gamma}$, δs_1, $-\tilde{\gamma}'$ and $-\delta s_2$ form a closed curve in Σ. Consider the integral of ω over the infinitesimal surface bounded by this curve. This integral vanishes because at every point the surface is (to first order) parallel to $\tilde{\gamma}$, which is an integral line of ω. But $\omega = d\theta$ and therefore, by Stokes theorem, the integral of θ along the closed curve vanishes as well. The integral of $\theta = p_a dq^a$ along δs_1 and δs_2 is zero because q^a is constant along these segments. Therefore

$$\int_{\tilde{\gamma}} \theta + \int_{-\tilde{\gamma}'} \theta = 0, \tag{3.47}$$

or

$$\delta \int_{\tilde{\gamma}} \theta = 0, \tag{3.48}$$

for any variation in the class considered. This is precisely the variational principle stated in section 3.2.

Hamilton equations. Consider first the case $k = 1$. Motions are one-dimensional. Parametrize the curve with an arbitrary parameter τ. That is, describe a motion (in Ω) with the functions

$(q^a(\tau), p_a(\tau))$. These functions satisfy the Hamilton system

$$\begin{aligned} H(q^a, p_a) &= 0, \qquad (3.49)\\ \frac{dq^a(\tau)}{d\tau} &= N(\tau)\ v^a(q^a(\tau), p_a(\tau)),\\ \frac{dp_a(\tau)}{d\tau} &= N(\tau)\ f_a(q^a(\tau), p_a(\tau)) \qquad (3.50) \end{aligned}$$

where

$$v^a(q^a, p_a) = \frac{\partial H(q^a, p_a)}{\partial p_a}, \qquad f_a(q^a, p_a) = -\frac{\partial H(q^a, p_a)}{\partial q^a}. \qquad (3.51)$$

The function $N(\tau)$ is called the "Lapse function". It is arbitrary. Different choices of $N(\tau)$ determine different parameters τ along the motion. To obtain a monotonic parametrization we need $N(\tau) > 0$. A preferred parametrization can be obtained by taking $N(\tau) = 1$, that is, replacing (3.50-3.51) by the equations (written in the usual compact form)

$$\dot{q}^a = \frac{\partial H}{\partial p_a}, \qquad \dot{p}_a = -\frac{\partial H}{\partial q^a}, \qquad (3.52)$$

where the dot indicates derivative with respect to τ. This choice is called the Lapse=1 gauge. It is not preferred in a physical sense. In particular, different but physically equivalent hamiltonians H defining the same surface Σ determine different preferred parametrizations. But it is often the easiest to compute with.

If $k > 1$, the function H has components H^j, with $j = 1, \ldots, k$ and motions are k dimensional surfaces. We can parametrize a motion with k arbitrary parameters $\vec{\tau} = \{\tau_j\}$. Namely represent with the $2n$ functions $q^a(\vec{\tau}), p_a(\vec{\tau})$ of k parameters τ_j. These equations satisfy the system given by (3.49) and

$$\frac{\partial q^a(\vec{\tau})}{\partial \tau_j} = N_j(\vec{\tau})\ \frac{\partial H^j(q^a, p_a)}{\partial p_a}, \qquad \frac{\partial p_a(\vec{\tau})}{\partial \tau_j} = -N_j(\vec{\tau})\ \frac{\partial H^j(q^a, p_a)}{\partial q^a}. \qquad (3.53)$$

A motion is determined by the full k dimensional surface in $\mathcal{C}$, but we can choose a particular curve $\vec{\tau}(\tau)$ on this surface, where τ is an arbitrary parameter, and represent the motion by the one-dimensional curve $q^a(\tau) = q^a(\vec{\tau}(\tau))$ in $\mathcal{C}$. This satisfies the system formed by (3.49) and

$$\frac{dq^a(\tau)}{d\tau} = N_j(\tau)\ \frac{\partial H^j(q^a, p_a)}{\partial p_a}, \qquad \frac{dp_a(\tau)}{d\tau} = -N_j(\tau)\ \frac{\partial H^j(q^a, p_a)}{\partial q^a}. \qquad (3.54)$$

for k arbitrary functions of one variable $N_j(\tau)$. Different choices of the functions $N_j(\tau)$ determine different curves on the single surface that defines a motion. These are gauge equivalent representations of the same motion.

It is important to stress that the parameters τ or τ_j are an artifact of this technique. They have no physical significance. They are absent in the geometric formalism as well as in the Hamilton-Jacobi formalism, as we shall see below. The physical content of the theory is in the motion in $\mathcal{C}$, not in the way the motion is parametrized. That is, the physical information is not in the functions $q^a(\tau)$: it is in the image of these functions in $\mathcal{C}$.

Relation with the variational principle. Parametrize the curve $\tilde{\gamma}$ with a parameter τ. The action (3.38) reads

$$S = \int d\tau\ p_a(\tau)\ \frac{dq^a(\tau)}{d\tau} \qquad (3.55)$$

The constraint (3.39) can be implemented in the action with lagrangian multipliers $N_i(\tau)$. This defines the action

$$S = \int d\tau \left(p_a\ \frac{dq^a}{d\tau} + N_i\ H^i(p_a, q^a) \right). \qquad (3.56)$$

Varying this action with respect to $N_i(\tau)$, $q^a(\tau)$ and $p_a(\tau)$ gives the Hamilton equation (3.49, 3.54).

Example: double pendulum. Consider the system defined by the Hamiltonian (3.40). The Hamilton equations (3.49, 3.52) in the Lapse=1 gauge give

$$\dot{a} = p_a, \qquad \dot{b} = p_b, \qquad \dot{p}_a = -a, \qquad \dot{p}_b = -b, \qquad a^2 + b^2 + p_a^2 + p_b^2 = E^2, \tag{3.57}$$

The general solution is

$$a(\tau) = A_a \sin(\tau), \qquad b(\tau) = A_b \sin(\tau + \beta). \tag{3.58}$$

where $A_a = E\sin\alpha$ and $A_b = E\cos\alpha$. The motions are given by the image in $\mathcal{C}$ of these curves. These are the ellipses (3.41). The parametrization of the curves (3.58) has no physical significance. The physics is in the unparametrized ellipses in $\mathcal{C}$ and in the relation between a and b they determine.

Hamilton-Jacobi. Hamilton-Jacobi formalism is elegant, general and powerful; it has a direct connection with quantum theory, and is conceptually clear. The *relativistic* formulation of Hamilton-Jacobi theory is simpler than the conventional nonrelativistic version, indicating that the relativistic formulation unveils a natural and general structure of mechanical systems.

The relativistic Hamilton-Jacobi formalism is given by the system of k partial differential equations

$$H\left(q^a, \frac{\partial S(q^a)}{\partial q^a}\right) = 0. \tag{3.59}$$

for the function $S(q^a)$ defined on the extended configuration space $\mathcal{C}$. Let $S(q^a, Q^i)$ be a family of solutions, parametrized by the $n-k$ constants of integration Q^i. Pose

$$f^i(q^a, P_i, Q^i) \equiv \frac{\partial S(q^a, Q^i)}{\partial Q^i} - P_i = 0 \tag{3.60}$$

for $n-k$ arbitrary constants P_i. This is the evolution equation. The constants Q^i, P_i coordinatize thus a $2(n-k)$ dimensional phase space Γ. This is the phase space.

The form of the relativistic Hamilton-Jacobi equation (3.59) is simpler than the usual nonrelativistic Hamilton-Jacobi equation (3.17). Furthermore, there is no equation to invert, as in the nonrelativistic formalism. Notice also that the function $S(q^a, Q^i)$ can be identified with the *principal* Hamilton-Jacobi function $S(t, q^i, Q^i) = Et + W(q^i, Q^i)$ of the nonrelativistic formalism, as well as with the *characteristic* Hamilton-Jacobi function $W(q^i, Q^i)$, since (3.59) is formally like (3.20) with vanishing energy. The two functions are in fact identified in the relativistic formalism.

Example: double pendulum. The Hamilton-Jacobi equation of the timeless system (3.40) is

$$\left(\frac{\partial S(a,b)}{\partial a}\right)^2 + \left(\frac{\partial S(a,b)}{\partial b}\right)^2 + a^2 + b^2 - 2E = 0. \tag{3.61}$$

A one parameter family of solution is given by

$$\begin{aligned} S(a,b,A) \;=\; & \frac{a}{2}\sqrt{A^2-a^2} + \frac{A^2}{2}\arctan\left(\frac{a}{\sqrt{A^2-a^2}}\right) \\ & + \frac{b}{2}\sqrt{2E-A^2-b^2} + \frac{2E-A^2}{2}\arctan\left(\frac{b}{\sqrt{2E-A^2-b^2}}\right). \end{aligned} \tag{3.62}$$

The general solution (3.41) of the system is directly obtained by writing

$$\frac{\partial S(a,b,A)}{\partial A} - \phi = 0. \tag{3.63}$$

Derivation of the Hamilton-Jacobi formalism. Since the phase space Γ is a symplectic space, locally we can choose canonical coordinates (Q^i, P_i) over it. These coordinates can be pulled back to Σ, where they are constant along the orbits. In facts, they label the orbits. Let $\theta_{ph} = P_i dQ^i$. Therefore $d\theta_{ph} = \omega$. But $\omega = d\theta = d(p_a dq^a)$. Thus on Σ we have

$$d(\theta_{ph} - \theta) = d(P_i dQ^i - p_a dq^a) = 0, \tag{3.64}$$

and therefore there should locally exist a function S on Σ such that

$$P_i dQ^i - p_a dq^a = dS \tag{3.65}$$

Let us choose q^a and Q^i as independent coordinates on Σ. Then the last equation reads

$$dS(q^a, Q^i) = p_a(q^a, Q^i)dq^a - P_i(q^a, Q^i)dQ^i, \tag{3.66}$$

that is

$$\frac{\partial S(q^a, Q^i)}{\partial q^a} = p_a(q^a, Q^i), \tag{3.67}$$

$$\frac{\partial S(q^a, Q^i)}{\partial Q^i} = -P_i(q^a, Q^i). \tag{3.68}$$

By the definition of Σ, we have $H(q^a, p_a) = 0$, which, using (3.67), gives the Hamilton-Jacobi equation (3.59). Equation (3.68) is then immediately the evolution equation (3.60).

In other words, $S(q^a, Q^i)$ is the generating function of a canonical transformation that relates the observables and their momenta (q^a, p_a) to new canonical variables (Q^i, P_i) satisfying $\dot{Q}^i = 0, \dot{P}_i = 0$. These new variables are constants of motions and therefore define Γ. The relation between $\mathcal{C}$ and Γ given by the canonical transformation equation (3.67–3.68) is the evolution equation.

3.2.3 Nonrelativistic systems as a special case

Here I discuss in more detail how the notions and the structures of conventional mechanics described in sec 3.1 are recovered from the relativistic formalism. A nonrelativistic system is simply a relativistic dynamical system in which one of the partial observables q^a is denoted t and called "time", and the hamiltonian H has the form

$$H = p_t + H_0 \tag{3.69}$$

where H_0 is independent from p_t and is called the nonrelativistic hamiltonian. The quantity $E = -p_t$ is called energy. The device that measures the partial observable t is called a clock.

The relativistic configuration space has therefore the structure

$$\mathcal{C} = R \times \mathcal{C}_0, \tag{3.70}$$

with coordinates $q^a = (t, q^i)$, where $i = 1, \ldots n$–1. The space $\mathcal{C}_0$ is the usual nonrelativistic configuration space. Accordingly, the cotangent space $\Omega = T^*\mathcal{C}$ has coordinates $(q^a, p_a) = (t, q^i, p_t, p_i)$.

If H has the form (3.69), the relativistic Hamilton-Jacobi equation (3.59) becomes the conventional nonrelativistic Hamilton-Jacobi equation (3.17).

Given a state and a value t of the clock observable, we can ask what are the possible values of the observables q^i such that (q^i, t) is a possible event. That is, we can ask what is the value of q^i "when" the time is t. The solution is obtained by solving the evolution function $f^i(q^i, t; Q^i, P_i) = 0$, for the q^i. This gives

$$q^i = q^i(t; Q^i, P_i). \tag{3.71}$$

which is interpreted as the evolution equation of the variables q^i in the time t. The form (3.69) of the hamiltonian guarantees that we can solve f with respect to the q^i's, because the Hamilton equation for t (in the Lapse=1 gauge) is simply $t = \tau$, which can be inverted.

In the parametrized hamiltonian formalism, the evolution equation for $t(\tau)$ is trivial and gives, taking advantage of the freedom in rescaling τ, just $t = \tau$. Using this, equations (3.53) become the conventional Hamilton equations and (3.49) simply fixes the value of p_t, namely of the energy.

In the presymplectic formalism, the surface Σ turns out to be

$$\Sigma = R \times \Gamma_0 \tag{3.72}$$

where the coordinate on R is the time t and $\Gamma_0 = T^*\mathcal{C}_0$ is the nonrelativistic phase space. The restriction of $\tilde{\theta}$ to this surface has the Cartan form

$$\theta = p_i dq^i - H_0 dt = \theta_0 - H_0 dt. \tag{3.73}$$

We can take the vector field X to have the form

$$X = \frac{\partial}{\partial t} + X_0 \tag{3.74}$$

where X_0 is a vector field on Γ_0. Then the equation of motion (3.46) reduces to the equation

$$(d\theta_0)(X_0) = -dH_0, \tag{3.75}$$

which is the geometric form of the conventional Hamilton equations. Thus, H determines how the variables in Γ_0 are correlated to the variable t. That is "how the variables in Γ_0 evolve in time". In this sense, the nonrelativistic hamiltonian H_0 generates "evolution in the time t". This evolution is generated in Γ_0 by the hamiltonian flow X_0 of H_0. A point $s = (q^i, p_i)$ in Γ_0 is taken to the point $s(t) = (q^i(t), p_i(t))$ where

$$\frac{ds(t)}{dt} = X_0(s(t)). \tag{3.76}$$

The evolution of any observable is defined by $A_t(s) = A(s(t)) = A(s,t)$ can be written, introducing the Poisson bracket notation

$$\{A, B\} = -X_A(B) = X_B(A) = \sum_i \left(\frac{\partial A}{\partial q^i}\frac{\partial B}{\partial p_i} - \frac{\partial A}{\partial p_i}\frac{\partial B}{\partial q^i} \right), \tag{3.77}$$

as

$$\frac{dA_t}{dt} = \{A_t, H_0\}. \tag{3.78}$$

Instantaneous states and relativistic states. The nonrelativistic definition of *state* refers to the properties of a system *at a certain moment of time*. Denote this conventional notion of state as the "instantaneous state". The space of the instantaneous states is the conventional nonrelativistic phase space Γ_0. Let's fix the value $t = t_0$ of the time variable, and characterize the instantaneous state in terms of the initial data. For the pendulum these are position and momentum (α_0, p_0), at $t = t_0$. Thus (α_0, p_0) are coordinates on Γ_0.

On the other hand, a relativistic state is a solution of the equation of motion. (If there is gauge invariance, a state is a gauge equivalence class of solutions of the equations of motion). The relativistic phase space Γ is the space of the solutions of the equations of motion.

Given a value t_0 of the time, there is a one-to-one correspondence between initial data and solutions of the equations of motion: Each solution of the equation of motion determines initial data at $t = t_0$; and each choice of initial data at t_0 determines uniquely a solution of the equations of motion. Therefore there is a one-to-one correspondence between instantaneous states and relativistic states. Therefore the relativistic phase space Γ is isomorphic to the nonrelativistic phase space: $\Gamma \sim \Gamma_0$. However, the isomorphism depends on the time t_0 chosen, and the physical interpretation of the two spaces is quite different. One is a space of states at a given time, the other a space of motions.

In the case of the pendulum, the nonrelativistic phase space, Γ_0, can be coordinatized with (α_0, p_0); the relativistic phase space Γ with (A, ϕ). The identification map $(A, \phi) \mapsto (\alpha_0,\ p_0)$ is given by

$$\alpha_0(A, \phi) = A \sin(\omega t_0 + \phi), \tag{3.79}$$

$$p_0(A, \phi) = \omega m A \cos(\omega t_0 + \phi). \tag{3.80}$$

The nonrelativistic phase space Γ_0 plays a double role in nonrelativistic hamiltonian mechanics: it is the space of the instantaneous states, but it is also the arena of nonrelativistic hamiltonian mechanics, over which H_0 is defined. In the relativistic context, this double role is lost: one must distinguish the cotangent space $\Omega = T^*\mathcal{C}$ over which H is defined, from the phase space Γ, which is the space of the motions. This distinction will become important in field theory, where Ω is finite dimensional, while Γ is infinite dimensional.

In a nonrelativistic system, X_0 generates a one-parameter group of transformation in Γ_0, the hamiltonian flow of H_0 on Γ_0. Instead of having the observables in $\mathcal{C}_0$ depending on t, one can shift perspective and view the observables in $\mathcal{C}_0$ as time independent objects and the states in Γ_0 as time dependent objects. This is a classical analog of the shift from the Heisenberg to the Schrödinger picture in quantum theory, and can be called the "classical Schrödinger picture".

In the relativistic theory there is no special "time" variable, $\mathcal{C}$ does not split naturally as $\mathcal{C} = R \times \mathcal{C}_0$, the constraints do not have the form $H = p_t + H_0$ and the description of the correlations in terms of "how the variables in $\mathcal{C}_0$ evolve in time" is not available in general. In a system that does not admit a nonrelativistic formulation, the classical Schrödinger picture, in which states evolve in time, is not available: only the relativistic notions of state and observable make sense.

Special relativistic systems. There are relativistic systems that do not admit a nonrelativistic formulation, as the example of the double pendulum discussed above. There are also systems that can be given a nonrelativistic formulation, but their structure is far more clean in the relativistic formalism. Lorentz invariant systems are typical examples. They can be formulated in the conventional hamiltonian picture only at the price of breaking Lorentz invariance. The choice of a preferred Lorentz frame specifies a preferred Lorentz time variable $t = x^0$. The predictions of the theory are Lorentz invariant, but the formalism is not. This way of dealing with the mechanics of a special relativistic systems hides the simplicity and symmetry of its hamiltonian structure. The relativistic hamiltonian formalism, exemplified below for the case of a free particle, on the other hand, is manifestly Lorentz invariant.

Example: Relativistic particle. The configuration space $\mathcal{C}$ is Minkowski space $\mathcal{M}$, with coordinates x^μ. The dynamics is given by the hamiltonian $H = p^\mu p_\mu + m^2$, which defines the mass-m Lorentz hyperboloid $\mathcal{K}_m$. The constraint surface Σ is therefore given by $\Sigma = T^*\mathcal{M}|_{H=0} = \mathcal{M} \times \mathcal{K}_m$. The null vectors of the restriction of $d\theta = dp_\mu \wedge dx^\mu$ to Σ are

$$X = p_\mu \frac{\partial}{\partial x^\mu}, \tag{3.81}$$

because $\omega(X) = p^\mu dp_\mu = 2d(p^2) = 0$ on $p^\mu p_\mu = -m^2$. The integral lines of X, namely the lines whose tangent is X are

$$x^\mu(\tau) = P^\mu \tau + X^\mu, \qquad p^\mu(\tau) = P^\mu \tag{3.82}$$

which give the physical motions of the particle. The space of these lines is six dimensional (It is coordinatized by the eight numbers (X^μ, P^μ) but $P^\mu P_\mu = -m^2$ and (P^μ, X^μ) defines the same line as $(P^\mu, X^\mu + P^\mu a)$ for any a), and represents the phase space. The motions are thus the timelike straight lines in $\mathcal{M}$.

Notice that all notions used are completely Lorentz invariant. A state is a timelike geodesic; an observable is any Minkowski coordinate, a correlation is a point in Minkowski space. The theory is about correlations between Minkowski coordinates, that is, observations of the particle at certain spacetime points. On the other hand, the split $\mathcal{M} = \mathbb{R} \times Q$ necessary to define the usual hamiltonian formalism, is observer dependent.

The relativistic formulation of mechanics is not only more general, but also more simple and elegant, and better operationally founded, than the conventional nonrelativistic formulation. This is true whether one uses the Hamilton equations, the geometric language, or the Hamilton-Jacobi formalism.

3.2.4 Discussion: Mechanics is about relations between observables

The key difference between the relativistic formulation of mechanics discussed in this chapter and the conventional one –and in particular between the relativistic definition of state and observable and the conventional one– is the role played by time. In the nonrelativistic context, time is a primary concept. Mechanics is defined as the theory of the evolution in time. In the definition considered here, on the other hand, no special partial observable is singled out as the independent variable. Mechanics is defined as the theory of the correlations between partial observables.

Technically, $\mathcal{C}$ does not split naturally as $\mathcal{C} = \mathbb{R} \times \mathcal{C}_0$, the constraints do not have the form $H = p_t + H_0$ and the Schrödinger-like description of correlations in terms of "how states and observables evolve in time" is not available in general.

It is important to understand clearly the meaning of this shift of perspective.

The first point is that *it is possible* to formulate conventional mechanics in this time-independent language. In fact, the formalism of mechanics becomes even more clean and symmetric (for instance, Lorentz covariant) in this language. This is a remarkable fact by itself. What is remarkable is that the formal structure of mechanics doesn't really treat the time variable on a different footing than the other variables. The structure of mechanics is the formalization of what we have understood about the physical structure of the world. Therefore, we can say that the physical (more precisely, mechanical) structure of the world is quite blind to the fact that there is anything "special" about the variable t.

Historically, the idea that in a relativistic context we need the time-independent notion of state has been advocated particularly by Dirac [135] and by Souriau [98]. The advantages of the relativistic notion of state are multifold. In special relativity, for instance, time transforms with other variables, and there is no covariant definition of instantaneous state. In a Lorentz invariant field theory, in particular, the notion of instantaneous state breaks explicit Lorentz covariance: the instantaneous state is the value of the field on a simultaneity surface, which is such for a certain observer only. The relativistic notion of state, on the other hand, is Lorentz invariant.

The second point is that this shift in perspective, however, is *forced* in general relativity, where the notion of a special spacelike surface over which initial data are fixed conflicts with diffeomorphism invariance. A generally covariant notion of instantaneous state, or a generally covariant notion of observable "at a given time", makes little physical sense. Indeed, none of the various notions of time that appear in general relativity (coordinate time, proper time, clock time) play the role that t plays in nonrelativistic mechanics. A consistent definition of state and observable in a generally covariant context cannot explicitly involve time.

The physical reason for this difference is discussed in chapter 2. In nonrelativistic physics, time and position are defined with respect to a system of reference bodies and clocks that are implicitly assumed to exist and not to interact with the physical system studied. In gravitational physics, one discovers that no body or clock exists which does not interact with the gravitational field: the gravitational field affects directly the motion and the rate of any reference body or clock. Therefore one cannot separate reference bodies and clocks from the dynamical variables of the system. General relativity –in fact, any general covariant theory– is always a theory of interacting variables that necessarily include the physical bodies and clocks used as references to characterize spacetime points. In the example of the pendulum discussed in section 3.2.1, for instance, we can assume that the pendulum itself and the clock do not interact. In a general relativistic context the two always interact, and $\mathcal{C}$ does not split into $\mathcal{C}_0$ and $\mathbb{R}$.

Summarizing, it is only in the nonrelativistic limit that mechanics can be seen as the theory of the evolution of the physical variables in time. In a fully relativistic context, *mechanics is a theory of correlations between partial observables.*

3.2.5 Space of boundary data $\mathcal{G}$ and Hamilton function S

I describe here the relativistic version of a structure that plays an important role in the quantum theory.

Hamilton function. Notice that the Hamilton function defined in (3.21) is naturally a function on (two copies of) the relativistic configuration space $\mathcal{C}$. In fact, its definition extends to the relativistic context: given two events q^a and q^a_0 in $\mathcal{C}$, the hamilton function is defined as

$$S(q^a, q^a_0) = \int_{\tilde{\gamma}} \theta. \tag{3.83}$$

where $\tilde{\gamma}$ is the orbit in Σ of the motion that goes from q^a_0 to q^a. This is also the value of the action along this motion. For instance, for a nonrelativistic system we can write

$$\begin{aligned} S(q^a, q^a_0) &= \int_\gamma \theta = \int_\gamma p_a dq^a \qquad (3.84)\\ &= \int_0^1 p_a(\tau)\dot{q}^a(\tau)d\tau = \int_0^1 \left(p_i(\tau)\dot{q}^i(\tau) + p_t(\tau)\dot{t}(\tau)\right) d\tau \\ &= \int_0^1 \left(p_i(\tau)\dot{q}^i(\tau) - H_0(\tau)\dot{t}(\tau)\right) d\tau \\ &= \int_{t_0}^t \left(p_i(t)\frac{dq^i(t)}{dt} - H_0(t)\right) dt \\ &= \int_{t_0}^t L\left(q^i, \frac{dq^i(t)}{dt}\right) dt, \qquad (3.85) \end{aligned}$$

where L is the Lagrangian. From the definition, we have

$$\frac{\partial S(q^a, q^a_0)}{\partial q^a} = p_a(q^a, q^a_0) \tag{3.86}$$

where $p_a(q^a, q^a_0)$ is the value of the momenta at the final event. Notice that this value depends on q^a *as well as on* q^a_0. The derivation of this equation is less obvious than what it looks at first sight: I leave the details to the acute reader.

It follows from (3.86) that $S(q^a, q^a_0)$ satisfies the Hamilton-Jacobi equation (3.59). The quantities q^a_0 can be seen as the Hamilton-Jacobi integration constants. Notice that they are n, not $n-1$. Equations (3.60) read now

$$f^a(q^a; q^a_0, p_{a0}) = \frac{\partial S(q^a, q^a_0)}{\partial q^a_0} + p_{a0} = 0. \tag{3.87}$$

Therefore the phase space is directly (over-)coordinatized by initial coordinates and momenta (q^a_0, p_{a0}). These are not independent for two reasons. First, they satisfy the equation $H = 0$. Second, different sets $(q^a_0(\tau), p_{a0}(\tau))$ along the same motion determine the same motion. Furthermore, one of the equations (3.87) turns out to be dependent from the others.

$S(q^a, q^a_0)$ satisfies the Hamilton-Jacobi equation in both sets of variables, namely it satisfies also

$$H\left(q^a_0, -\frac{\partial S(q^a, q^a_0)}{\partial q^a_0}\right) = 0, \tag{3.88}$$

where the minus sign comes from the fact that the second set of variables is in the lower integration boundary in (3.83).

If there is more than one physical motion γ connecting the boundary data, the Hamilton function is multivalued. If $\gamma_1, \ldots, \gamma_n$ are distinct solutions with the same boundary values, we denote its different branches as

$$S_i(q_1^a, q_2^a) = \int_{\tilde{\gamma}_i} \theta. \tag{3.89}$$

The Hamilton function is strictly related to the quantum theory. It is the phase of the propagator $W(q^a, q_0^a)$, which, as we shall see in chapter 5, is the main object of the quantum theory. If S is single valued, we have

$$W(q^a, q_0^a) \sim A(q^a, q_0^a)\ e^{\frac{i}{\hbar} S(q^a, q_0^a)} \tag{3.90}$$

up to higher terms in $\hbar$. If S is multivalued,

$$W(q^a, q_0^a) \sim \sum_i A_i(q^a, q_0^a)\ e^{\frac{i}{\hbar} S_i(q^a, q_0^a)}. \tag{3.91}$$

Example: free particle. In the case of the free particle, the value of the classical action, along the motion is

$$\begin{aligned} S(x,t,x_0,t_0) &= \int_0^1 (p_t \dot{t} + p\dot{x}) = p_t \int_{t_O}^{t} dt + p \int_{x_0}^{x} dx \\ &= -\frac{m(x-x_0)^2}{2(t-t_0)} + m\frac{(x-x_0)^2}{t-t_0} \\ &= \frac{m(x-x_0)^2}{2(t-t_0)}. \end{aligned} \tag{3.92}$$

It is easy to check that S solves the Hamilton-Jacobi equation of the free particle. The first of the two equations (3.87) gives the evolution equation

$$\frac{\partial S(x,t,x_0,t_0)}{\partial x_0} + p_0 = -m\frac{x-x_0}{t-t_0} + p_0 = 0. \tag{3.93}$$

The second equation constrains the p_t integration constant

$$\frac{\partial S(x,t,x_0,t_0)}{\partial t_0} + p_{t0} = -\frac{1}{2m}p_0^2 + p_{t0} = 0. \tag{3.94}$$

Recall that the propagator of the Schrödinger equation of the free particle is

$$W(x,t,x_0,t_0) = \frac{1}{\sqrt{i\hbar(t-t_0)}}\ e^{\frac{i}{\hbar}\frac{m(x-x_0)^2}{2(t-t_0)}} = \frac{1}{\sqrt{i\hbar(t-t_0)}}\ e^{\frac{i}{\hbar}S(x,t,x_0,t_0)}. \tag{3.95}$$

Example: double pendulum. The Hamilton function of the timeless system (3.40) can be computed directly from its definition. This gives

$$S(a,b,a',b') = S\Big(a,b,a',b';\ A(a,b,a',b')\Big) \tag{3.96}$$

where

$$S(a,b,a',b';\ A) = S(a,b,A) - S(a',b',A), \tag{3.97}$$

$S(a,b,A)$ is given in (3.62) and $A(a,b,a',b')$ is the value of A of the ellipses (3.41) that crosses (a,b) and (a',b'). This value can be obtained by noticing that (3.58) imply, with little algebra, that

$$A^2 = \frac{a^2 + a'^2 - 2aa'\cos\tau}{\sin^2\tau} \tag{3.98}$$

and

$$M = \frac{(a^2 + b^2 + a'^2 + b'^2) - 2(aa' + bb')\cos\tau}{\sin^2\tau} \tag{3.99}$$

The second equation can be solved for $\tau(a,b,a',b')$, inserting this in the first gives $A(a,b,a',b')$. It is not complicated to check that the derivative of $\partial S(a,b,a',b';\ A)/\partial A$ vanishes when $A = A(a,b,a',b')$, and using this it is easy to see that (3.96) solves the Hamilton-Jacobi equation in both set of variables.

Notice that, for given (a, b, a', b') equation (3.98) gives A as a function $A(\tau)$ of τ. We can therefore consider also the function

$$S(a, b, a', b';\ \tau) = S(a, b, A(\tau)) - S(a', b', A(\tau)), \tag{3.100}$$

which is the value of the action of the nonrelativistic system formed by two harmonic oscillators evolving in a physical time τ, with a nonrelativistic hamiltonian H, that is, it is the Hamilton function of this system. With some algebra, this can be written also as

$$S(a, b, a', b';\ \tau) = M\tau + \frac{(a^2 + b^2 + a'^2 + b'^2)\cos\tau - 2(aa' + bb')}{\sin\tau}. \tag{3.101}$$

As for A, we have immediately

$$\left.\frac{\partial S(a, b, a', b';\ \tau)}{\partial \tau}\right|_{\tau=\tau(a,b,a',b')} = 0 \tag{3.102}$$

This means that the Hamilton function of the timeless system is numerically equal to the Hamilton function of the two oscillators for the "correct" time τ needed to go from (a', b') to (a, b) staying on a motion of total energy E. And that this "correct" time $\tau = \tau(a, b, a', b')$ is the one that minimizes the Hamilton function of the two oscillators.

More precisely, for given (a, b, a', b') there are two path connecting (a', b') with (a, b): these are the two paths in which the ellipses that goes through (a', b') and (a, b) is cut by these two points. Denote S_1 and S_2 the two values of the action along the these paths. Their relation is easily obtained by noticing that the action along the entire ellipses is easily computed as

$$S_1 + S_2 = 2\pi E. \tag{3.103}$$

The space of the boundary data $\mathcal{G}$. The Hamilton function is a function on the space $\mathcal{G} = \mathcal{C} \times \mathcal{C}$. An element $\alpha \in \mathcal{G}$ is an ordered pair of elements of the extended configuration space $\mathcal{C}$: $\alpha = (q^a, q_0^a)$. Notice that α is the ensemble of the boundary conditions for a physical motion. For a nonrelativistic system, $\alpha = (t, q^i, t_0, q_0^i)$; the motion begins at q_0^i at time t_0 and ends at q^i at time t.

The space $\mathcal{G}$ carries a natural symplectic structure. In fact, let $i : \mathcal{G} \to \Gamma$ be the map that sends each pair to the orbit that the pair defines. Then we can define the two form $\omega_{\mathcal{G}} = i^* \omega_{ph}$, where ω_{ph} is the symplectic form of the phase space, defined in section 3.2.2. In other words, $\alpha = (q^a, q_0^a)$ can be taken as a natural over-coordinatization of the phase space. Instead of coordinatizing a motion with initial positions and momenta, we coordinatize it with initial and final positions. In these coordinates, the symplectic form is given by $\omega_{\mathcal{G}}$.

The two form $\omega_{\mathcal{G}}$ can be computed without having first to compute Γ and ω_{ph}. Denote $\tilde{\gamma}_\alpha$ the orbit in Σ with boundary data α, and γ_α its projection to $\mathcal{C}$. Then α is the boundary of γ_α. We write $\alpha = \partial\gamma_\alpha$. Denote s and s_0 the initial and final points of $\tilde{\gamma}_\alpha$ in Σ. That is, $s = (q^a, p_a)$ and $s_0 = (q_0^a, p_{0a})$, where in general both p_a and p_{0a} depend on q^a and on q_0^a. Let $\delta\alpha = (\delta q^a, \delta q_0^a)$ be a vector (an infinitesimal displacement) at α. Then the following is true:

$$\begin{aligned} \omega_{\mathcal{G}}(\alpha)(\delta_1\alpha, \delta_2\alpha) &= \omega_{\mathcal{G}}(q^a, q_0^a)((\delta_1 q^a, \delta_1 q_0^a), (\delta_2 q^a, \delta_2 q_0^a)) \\ &= \omega(s)(\delta_1 s, \delta_2 s) - \omega(s_0)(\delta_1 s_0, \delta_2 s_0). \end{aligned} \tag{3.104}$$

Notice that $\delta_1 s$, the variation of s, is determined by $\delta_1 q$ as well as by $\delta_1 q_0$, and so on. This equation expresses $\omega_{\mathcal{G}}$ directly in terms of ω. As we shall see, this equation admits an immediate generalization in the field theoretical framework, where ω will be a five-form, but $\omega_{\mathcal{G}}$ is a two-form.

Now fix a pair $\alpha = (q^a, q_0^a)$ and consider a small variation of only one of its elements. Say

$$\delta\alpha = (\delta q^a, 0). \tag{3.105}$$

This defines a vector $\delta\alpha$ at α on $\mathcal{G}$, which can be pushed forward to Γ. If the variation is along the direction of the motion, then the push forward vanishes, that is $i_* \delta\alpha = 0$, because α and

$\alpha + \delta\alpha$ define the same motion. It follows that if the variation is along the direction of the motion, $\omega_{\mathcal{G}}(\delta\alpha) = 0$. Therefore the equation

$$\omega_{\mathcal{G}}(X) = 0. \tag{3.106}$$

gives the solutions of the equations of motion.

Thus, the pair $(\mathcal{G}, \omega_{\mathcal{G}})$ contains all the relevant information on the system. The null directions of $\omega_{\mathcal{G}}$ define the physical motions, and if we divide $\mathcal{G}$ by these null directions, the factor space is the physical phase space, equipped with the physical symplectic structure.

Example: free particle. The space $\mathcal{G}$ has coordinates $\alpha = (t, x, t_0, x_0)$. Given this point in $\mathcal{G}$, there is one motion that goes from (t_0, x_0) to (t, x), which is

$$t(\tau) = t_0 + (t - t_0)\tau, \tag{3.107}$$
$$x(\tau) = x_0 + (x - x_0)\tau. \tag{3.108}$$

Along this motion,

$$p = m\frac{x - x_0}{t - t_0}, \tag{3.109}$$
$$p_t = -\frac{(x - x_0)^2}{2m(t - t_0)^2}. \tag{3.110}$$

The map $i : \mathcal{G} \to \Gamma$ is thus given by

$$P = p = m\frac{x - x_0}{t - t_0}, \tag{3.111}$$
$$Q = x - \frac{p}{m}t = x - \frac{x - x_0}{t - t_0}t, \tag{3.112}$$

and therefore the two-form $\omega_{\mathcal{G}}$ is

$$\begin{aligned}\omega_{\mathcal{G}} &= i^*\omega_\Gamma = dP(t, x, t_0, x_0) \wedge dQ(t, x, t_0, x_0) \\ &= m\, d\frac{x - x_0}{t - t_0} \wedge d\left(x - \frac{x - x_0}{t - t_0}t\right) \\ &= \frac{m}{t - t_0}\left(dx - \frac{x - x_0}{t - t_0}dt\right) \wedge \left(dx_0 - \frac{x - x_0}{t - t_0}dt_0\right).\end{aligned} \tag{3.113}$$

It is immediate to see that a variation $\delta\alpha = (\delta t, \delta x, 0, 0)$ (at constant (x_0, t_0)) such that $\omega_{\mathcal{G}}(\delta\alpha) = 0$ must satisfy

$$\delta x = \frac{x - x_0}{t - t_0}\,\delta t. \tag{3.114}$$

This is precisely a variation of x and t along the physical motion (determined by (x_0, t_0)). Therefore $\omega_{\mathcal{G}}(\delta\alpha) = 0$ gives again the equations of motion. The two null directions of $\omega_{\mathcal{G}}$ are thus given by the two vector fields

$$X = \frac{x - x_0}{t - t_0}\partial_x + \partial_t, \tag{3.115}$$
$$X_0 = \frac{x - x_0}{t - t_0}\partial_{x_0} + \partial_{t_0}, \tag{3.116}$$

which are in involution (their Lie bracket vanishes), and therefore define a foliation of $\mathcal{G}$ with two-dimensional surfaces. These surfaces are parametrized by P and Q, given in (3.111,3.112), and in fact

$$X(P) = X(Q) = X_0(P) = X_0(Q) = 0. \tag{3.117}$$

In fact, we have simply recovered in this way the physical phase space: the space of these surfaces is the phase space Γ and the restriction of $\omega_{\mathcal{G}}$ to it is the physical symplectic form ω_{ph}.

Physical predictions from S. There are several different was of deriving physical predictions from the Hamilton function $S(q^a, q_0^a)$.

- Equation (3.87) gives the evolution function f from the Hamilton function.

- If we can measure the partial observables q^a as well as their momenta p_a, then the Hamilton function can be used for making predictions as follows. Let

$$\begin{aligned} p_a^1(q_1^a, q_2^a) &= \frac{\partial S(q_1^a, q_2^a)}{\partial q_1^a}, \\ p_a^2(q_1^a, q_2^a) &= \frac{\partial S(q_1^a, q_2^a)}{\partial q_2^a}. \end{aligned} \tag{3.118}$$

  The two equations

$$\begin{aligned} p_a^1 &= p_a^1(q_1^a, q_2^a), \\ p_a^2 &= p_a^2(q_1^a, q_2^a) \end{aligned} \tag{3.119}$$

  relate the four partial observables of the quadruplet $(q_1^a, p_a^1, q_2^a, p_a^2)$. The theory predicts that it is possible to observe the quadruplet $(q_1^a, p_a^1, q_2^a, p_a^2)$ only if this satisfies (3.119). In this way the classical theory determines which combinations of values of partial observables can be observed.

- Alternatively, we can fix two points q_i^a and q_f^a in $\mathcal{C}$ and ask whether a third point q^a is on the motion determined by q_i^a and q_f^a. That is, ask wether or not we could observe the correlation q^a, given that the correlations q_i^a and q_f^a are observed. A moment of reflection will convince the reader that if the answer to this question is positive then

$$S(q_f^a, q^a) + S(q^a, q_i^a) = S(q_f^a, q_i^a), \tag{3.120}$$

  because the action is additive along the motion. Furthermore, the incoming momentum at q^a and the outgoing one must be equal, therefore

$$\frac{\partial S(q_f^a, q^a)}{\partial q^a} = -\frac{\partial S(q^a, q_f^a)}{\partial q^a}. \tag{3.121}$$

3.2.6 Evolution parameters

A physical system is often defined by an action which is the integral of a lagrangian in an evolution parameter. But there are two different physical meanings that the evolution parameter may have.

We have seen that the variational principle governing any hamiltonian system can be written in the form (here $k = 1$)

$$S = \int d\tau \left(p_a \frac{dq^a}{d\tau} + N\ H(p_a, q^a) \right). \tag{3.122}$$

The action is invariant under reparametrizations of the evolution parameter τ. The evolution parameter τ has no physical meaning: there is no measuring device related to it.

On the other hand, consider a nonrelativistic system, where $q^a = (t, q^i)$ and $H = p_t + H_0$. The action (3.122) becomes

$$S = \int d\tau \left(p_t \frac{dt}{d\tau} + p_i \frac{dq^i}{d\tau} + N\ (p_t + H_0(p_i, q^i)) \right). \tag{3.123}$$

Varying N we obtain the equation of motion

$$p_t = -H_0. \tag{3.124}$$