UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

Eduardo Amâncio Barbosa Oliveira

Vacuum energy in modern cosmology: an analysis of quantum field theory in curved spaces and its application to cosmological spacetimes

São Carlos

2022

Eduardo Amâncio Barbosa Oliveira

Vacuum energy in modern cosmology: an analysis of quantum field theory in curved spaces and its application to cosmological spacetimes

Dissertation presented to the Graduate Program in Physics at the Instituto de Física de São Carlos, Universidade de São Paulo to obtain the degree of Master of Science.

Concentration area: Theoretical and Experimental Physics

Advisor: Prof. Daniel Augusto Turolla Vanzella

Corrected version
(Original version available on the Program Unit)

São Carlos
2022

Oliveira, Eduardo Amâncio Barbosa
Vacuum energy in modern cosmology: an analysis of quantum field theory in curved spaces and its application to cosmological spacetimes / Eduardo Amâncio Barbosa Oliveira; advisor Daniel Augusto Turolla Vanzella - corrected version -- São Carlos 2022.
246 p.

Dissertation (Master's degree - Graduate Program in Theoretical and Experimental Physics) -- Instituto de Física de São Carlos, Universidade de São Paulo - Brasil , 2022.

1. Quantum field theory in curved spaces. 2. Vacuum energy. 3. Dark energy. 4. Inflation. 5. Particle creation. I. Vanzella, Daniel Augusto Turolla, advisor. II. Title.

ACKNOWLEDGEMENTS

Atravessar dois anos de mestrado em meio a uma pandemia que virou todas as nossas vidas do avesso foi uma experiência singular e extremamente desafiadora, e certamente não teria sido possível sem o suporte das muitas pessoas que conviveram e estiveram comigo nesse período tão peculiar. Deixo aqui os meus agradecimentos a algumas delas:

Aos meus pais, Amoacy e Ludmila, pelo apoio incondicional que me permitiu chegar aqui.

Ao meu orientador, Daniel Vanzella, por toda a física que me ensinou desde os anos de graduação, pela oportunidade de mestrado num tema tão interessante, pelas minunciosas discussões e observações, e pela desproporcional confiança depositada em mim ao longo desses últimos 2 anos. E ao meu ex-orientador, Emanuel Henn, pelos ensinamentos edificantes nos primeiros passos da minha vida acadêmica e por todo o apoio posterior.

Ao Japonês, pela grande ajuda com a formatação do trabalho, e pelas das cautelosas e mascaradas aventuras.

Ao Pestana e ao Hot Wheels, pelas longas pedaladas pelas ruas, trilhas e cachoeiras de São Carlos, que tanto contribuíram para o meu bem-estar e sanidade nesse período apocalíptico.

Ao Felipe (Neves), por infernizar a minha vida como ninguém.

À Giórgia, pelo carinho, suporte, e por todo o escopo de experiências e memórias que ajudaram a tornar esses turbulentos anos tão especiais.

E ao Momo (Buendía, Momão), meu onipresente companheiro de pandemia, com quem pude contar em todos os momentos de necessidade, num período em que necessidades não faltaram. Por toda a amizade e companheirismo, pelas irregulares jogatinas e escassas aventuras, e pelo suporte tecnológico e humano para produzir 13 das figuras que ilustram esse texto, meu muito obrigado.

Este trabalho foi financiado pelo Centro Nacional de Desenvolvimento Científico e Tecnológico, CNPq, no processo de número 131012/2020-7.

“Era ainda jovem demais para saber
que a memória do coração elimina as más lembranças e enaltece as boas
e que graças a esse artifício conseguimos suportar o passado.
Mas quando voltou a ver do convés do navio
o promontório branco do bairro colonial,
os urubus imóveis nos telhados,
a roupa dos pobres estendida a secar nas sacadas,
compreendeu até que ponto tinha sido uma vítima fácil das burlas caritativas da saudade.”

Gabriel García Marquez

ABSTRACT

OLIVEIRA, E. A. B. **Vacuum energy in modern cosmology:** an analysis of quantum field theory in curved spaces and its application to cosmological spacetimes. 2022. 246p. Dissertation (Master in Science) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2022.

The last decades have witnessed an unprecedented advancement in our knowledge of the large scale universe. In particular, increasingly accurate cosmological observations have allowed us to discover a form of "dark energy", which presently dominates the expansion of the universe – making it accelerated. On the other hand, fundamental problems in the standard (ΛCDM) cosmological model point towards the possibility of a primordial inflationary period. Both these expansion phases have in common the fact that they should be governed by forms of energy with properties much similar to those of vacuum energy of classical or quantum fields. In the meanwhile, quantum field theory in curved spaces (QFTCS) has proved a rich framework to analyze phenomena of a quantum nature in regimes where spacetime curvature is relevant, but not too extreme, and, particularly, it yields novel insights on the structure and dynamics of quantum vacuum. In this dissertation, we make a thorough exposition of the fundamentals of QFTCS and present some of its applications in cosmological spacetimes. Particular attention is given to the construction of an empirical notion of particles through an idealized model of particle detectors, and to the phenomenon of particle creation in expanding FLRW spacetimes. Further, we develop the procedure of adiabatic renormalization, and use it to compute the renormalized stress tensor in these spacetimes. For a noninteracting scalar field in exponentially expanding (de Sitter) spaces, we find that these results take the form of a cosmological constant, although a quantitatively self-consistent value with the background expansion can only be found at Planckian densities. We also present a construction of a simple inflationary model, driven by a self-interacting classical scalar field, and show how the quantized fluctuations of this field could give rise to a nearly scale-invariant power spectrum, like the one that is currently observed in the Cosmic Microwave Background.

Keywords: Quantum field theory in curved spaces. Vacuum energy. Dark energy. Inflation. Particle creation.

RESUMO

OLIVEIRA, E. A. B. **Energia de Vácuo na Cosmologia Moderna:** uma análise dos fundamentos de teoria quântica de campos em espaços curvos e suas aplicações a espaçostempos cosmológicos. 2022. 246p. Dissertação (Mestrado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2022.

As últimas décadas testemunharam um avanço sem precedentes no nosso conhecimento do universo em larga escala. Em particular, medidas cosmológicas cada vez mais precisas nos permitiram descobrir uma forma de "energia escura", que atualmente domina a expansão do universo – tornando-a acelerada. Por outro lado, problemas fundamentais no modelo cosmológico padrão (ΛCDM) apontam para a possibilidade de um período inflacionário primordial. Ambas essas fases de expansão têm em comum o fato de que elas deveriam ser governadas por formas de energia com propriedades muito similares àquelas da energia de vácuo de campos clássicos ou quânticos. Enquanto isso, teoria quântica de campos em espaços curvos (TQCEC) se mostrou um rico paradigma para analisar fenômenos de natureza quântica em regimes onde a curvatura do espaçotempo é relevante, mas não demasiado extrema, e, particularmente, ela provê novos *insights* sobre a estrutura e a dinâmica do vácuo quântico. Nesta dissertação, nós fazemos uma exposição detalhada dos fundamentos de TQCEC e apresentamos algumas das suas aplicações a espaçostempos cosmológicos. Particular atenção é dada à construção de uma noção empírica do conceito de partícula através de um modelo idealizado de detectores de partículas, e ao fenômeno de criação de partículas em espaçostempos de FLRW em expansão. Ademais, desenvolvemos aqui o procedimento de renormalização adiabática, e o usamos para computar o tensor energia-momentum renormalizado nesses espaçostempos. Para um campo escalar livre em espaços em expansão exponencial (espaços de de Sitter), encontramos resultados na forma de uma constante cosmológica; esta, todavia, só apresenta um valor quantitativamente autoconsistente com a expansão cósmica de fundo em escalas planckianas. Também apresentamos a construção de um modelo inflacionário simples, governado por um campo escalar clássico autointeragente, e mostramos como as flutuações quantizadas desse campo podem dar origem a um espectro aproximadamente invariante de escala, como o que é atualmente observado na Radiação Cósmica de Fundo.

Palavras-chave: Teoria quântica de campos em espaços curvos. Energia de vácuo. Energia escura. Inflação. Criação de partículas.

LIST OF SYMBOLS

$\hbar, c, G$ Fundamental physical constants. In this work, we employ natural units, $\hbar = c = 1$, in chapters 1-4, and Planck units, $\hbar = c = G = 1$, in chapter 5.

v^a, ω_a Spacetime vectors, dual vectors (covectors) and general tensors are written in the abstract index notation. Abstract indices will be latin letters ranging from a to h; vectors are denoted with a single upper index, dual vectors with a single lower index, and general tensors may have multiple upper and lower indexes. Exceptionally, we also use a single latin subscript to compactly denote multiple or composite fields: ϕ_a.

v^μ, ω_μ Greek indexes denote general vector/tensor components, each running from 0 to n, where n denotes the spacetime dimension. Latin indexes from i ownwards denote spatial components, running from 1 to n. Following the Einstein notation, repeated indices denote an implicit sum, unless otherwise stated.

g_{ab} Spacetime metric. In this work, we use the metric signature of Birrell & Davies (1): $(+, -, ...-)$. The flat Minkowski metric is denoted by η_{ab} and its components in a Global Inertial Frame are $\eta_{\mu\nu} = (1, -1, ... - 1)$.

μ The greek letter μ is often used to denote a measure on an arbitrary measure space, such as the spectrum of a linear operator H, $\sigma(H)$, or a manifold $\mathcal{M}$. For a metric manifold $(\mathcal{M}, g_{ab})$, $d\mu_g(x)$ will denote the natural volume element on $\mathcal{M}$ induced by g_{ab}.

$T_{(ab)}, T_{[ab]}$ Curly and square brackets are respectively used to denote complete symmetrization and antissymetrization over the encompassed indices. For example, $T_{(ab)} \equiv \frac{1}{2}(T_{ab} + T_{ba})$ and $T_{[ab]} \equiv \frac{1}{2}(T_{ab} - T_{ba})$.

$\pounds_v$ Lie derivative with respect to a vector field v^a.

$\mathbf{x}$, $\mathbf{y}$ Boldface letters denote vectors in $\mathbb{R}^n$, particularly, ordinary 3-dimensional spatial vectors. From chapter 3 onward, they are often used to compactly denote a set of spatial coordinates $\mathbf{x} \equiv \{x^i\}_{i=1,2,3}$ *even when spatial surfaces $\Sigma \subset \mathcal{M}$ do not possess a linear structure.*

CONTENTS

1 INTRODUCTION

The last 100 years have witnessed an unprecedented advancement in our understanding of the large scale universe. Einstein's theory of General Relativity – which radically changed the way we see spacetime, providing it with a dynamical character –, along with Hubble's first observations of departing galaxies (31), paved the way for the realisation (and eventual scientific consensus) that universe itself is in expansion. Already in 1920's and 30's emerged the so-called Friedman-Lemâitre-Robertson-Walker (FLRW) models[A]; based on the simplifying hypotheses of a spatially homogeneous and isotropic spacetime (which turns out to be quite accurate in very large scales) these models shaped much of the development of cosmology throughout the 20th century, laying the fundamentals for predictions about the universe's large scale structure and evolution based on a few of its average properties and parameters. But it was particularly in the last two or three decades that observation techniques and technologies were sufficiently developed to allow precise measurements of cosmological parameters – bringing uncertainties that were often of the same order (or grater!) than the measured values themselves down to just a few percentile points – and put tighter constraints in our models, allowing more rigorous consistency tests, as well as new, more precise and specific predictions.

Up to this date, the standard cosmological model (also known as the ΛCDM model) describes our cosmological observations with astounding precision[B], relying only in very few fundamental assumptions and parameters. However, in spite of being so observationally successful, this model suffers from many fundamental problems. These precise measurements of cosmological parameters allowed us to infer (assuming that General Relativity holds accurately in very large scales) much information about the matter and energy content of the universe. Astonishingly, the vast majority of the energy in the universe does not seem to be in the form of any known matter (which only seems to amount to about 5% of it), but rather in forms that we have only been able to detect gravitationally – and thus, particularly, that do not interact with light –, composing the so-called dark sector. This sector is divided in two major components: on relatively small (astrophysical) scales, the observed behaviour of massive matter structures (particularly, the rotation velocities of galaxies and peculiar velocities of galaxy clusters), as well as the formation of structure itself, requires for the presence of *Cold Dark Matter* (CDM), which composes around 25% of the total energy content; although exotic and not directly observable, dark matter seems to behave rather regularly gravitationally (in a much similar

[A] These were found independently and complementarily by many authors; for the seminal works of the 4 mentioned above, see (33–36).

[B] Although, more recently, interesting problems are starting to arise due to increasingly precise measurements, the most prominent of which is the H_0 tension. (44)

manner to ordinary baryonic matter). On very large (cosmological) scales, however, the expansion of the universe seems to be presently dominated by a much stranger energy form, the so-called *Dark Energy*, which comprises the remaining 70% ; it is not only undetectable through any nongravitational means, but it also (i) does not seem to form any types of structures, being distributed in a highly homogeneous way throughout the universe, and (ii) presents extremely negative pressures, with magnitude comparable to that of its energy density, which results in an effectively *repulsive* gravitational behaviour, causing the present cosmic expansion to be *accelerated*, rather than decelerated. Such exotic properties, although notably alien to most known physical systems, do appear naturally elsewhere: the vacuum energy of quantum (and classical) fields. As we shall see later in the present work, it is not unusual for renormalization procedures to yield negative expectation values of energy and/or pressure, even when these are classically positive-definite. Furthermore, it is a form of energy that permeates all space, so that it should be quite natural for it to be homogeneous and devoid of structure[C]. Finally, although there are numerous difficulties and indeterminacies in the calculation of renormalized energy, momentum and stress observables, the simplest form these quantities can take is precisely that of a cosmological constant Λ, which is exactly the form that Dark Energy seems to take.

To make its matters worse, the ΛCDM model is also full of ‘coincidences’ that are very difficult to explain from first principles. First and foremost, why is the universe so spatially homogeneous? And notably so in the past, before matter gravitationally collapsed and formed astrophysical structures; particularly, back in the Cosmic Microwave Background (CMB) formation the fluctuations in density and temperature were extremely small, with relative anysotropies of the order of $\mathcal{O}(10^{-5})$. In the standard Hot Big Bang scenario (in which the radiation era extends all the way back to a primordial singularity – the Big Bang) the patches of the sky that are causally connected should be no larger than about 2°. If widely separated portions of this early plasma have not had time to thermalize, and come to an equilibrium density, how come do they have such astoundingly similar temperatures and densities?[D] Furthermore, among all the possible values for spatial curvature in a homogeneous isotropic universe why is it so close (if not exactly equal) to 0? The matter becomes particularly acute when we note that any nonzero values of curvature (to which we can associate an effective energy density) tend to rapidly become dominant over ordinary matter components with nonnegative pressures[E]; if we were to

[C] This is indeed found to be the case for simple noninteracting fields *in homogeneous and isotropic spacetimes*, but, as we shall see throughout this dissertation, we should not underestimate how complex vacuum can be.

[D] Well, one could say they were just extremely homogeneous and uniform to start with. Although not impossible, this extreme level of *fine-tuning* in initial conditions makes for an arguably implausible and artificial explanation. We shall discuss the matter of initial conditions in further detail in section 5.3.

[E] Of course, this does not apply to dark energy. However, extremely small values of curvature adjust initial conditions at the Planck time $t_p \sim 10^{-43}s$, one would have to fine-tune the matter density to its so called critical value ρ_c in about 1 part in 10^{-41} so that universe would remain nearly flat up to the present time. One way to address all of these issues is by postulating a very brief primordial inflationary phase, which would have lasted about $\sim 10^{-33}s$ and during which the universe would have undergone an extremely fast and accelerated expansion, inflating by a factor of at least about 10^{26}. Such a wild proposal is, of course, extremely hard to probe and highly open to speculation. Nevertheless, it is quite widely accepted, since it provides a unified solution to many issues, as well as an arguably natural framework for studying the CMB fluctuation spectrum as due to primordial vacuum fluctuations of an inflation-driving field (so-called inflaton), stretched to wavelengths greater than the Hubble horizon during inflation. Not surprisingly, in order to bring about this primordial period of accelerated expansion, one needs a form of energy with peculiar properties much similar to those of dark energy, which could also be encompassed by vacuum energy.

In the meanwhile, quantum field theory in curved spacetimes (QFTCS) has proved a very rich and profound paradigm to analyze many intrinsically quantum phenomena in regimes where the spacetime curvature is relevant, but not too extreme (below Planck scales). In this approach, one avoids the (so far overwhelming) difficulties for obtaining a full quantum theory of matter and gravity, by quantizing only matter fields in classical curved background geometries. One of its most notable achievements is the prediction that Black Holes, rather than being perfectly opaque, actually emit thermal radiation – the well-known Hawking radiation (45) –, and can be meaningfully assigned with both temperature and entropy[F]. Equally noteworthy is the discovery of a flat space analogous to this thermal radiation, the Unruh effect (46). These effects turn out to provide deep insights in the nature of the quantum vacuum, and reveal that the concept of particles is considerably more malleable and observer-dependent than one could intuitively conceive.

In a cosmological context, the theory's novel features regarding vacuum energy provide a variety of theoretical possibilities to investigate dark energy, as well as the very early universe, particularly, a primordial inflationary period and some of its observational consequences today. The estimates that we can make for the necessary duration of inflation indicate that at least a significant portion of it should have occurred in extreme, but yet sub-Planckian regimes, where QFTCS is expected to hold[G]. In fact, one can find

in the past would have made it become dominant on time scales many orders of magnitude below than those for which DE became relevant.

[F] Although these quantities are already computable in the classical framework of GR (5) (up to a multiplicative constant in each, which is ultimately fixed by $\hbar$, but which disappears in the product TS), their physical meaning hardly seems to transcend a mere thermodynamical *analogy* before taking quantum effects in consideration.

[G] As we shall discuss in further detail in chapter 5, this restriction is somewhat tautological with our initial assumptions that our models are describable in terms of classical spacetimes, reasonable models for inflation whose average behaviour can be described in terms of the vacuum energy of classical fields. However, considering the quantization of these fields, one is able to find a considerably richer structure for this vacuum energy; among other things, it allows one to draw sensible predictions for the very small fluctuations that we observe in the CMB today.

In the present work, we attempt at providing a thorough and comprehensive exposition of QFTCS, particularly on vacuum energy and its applications in cosmological contexts. The dissertation is divided as follows: in Chapter 2, we lay some fundamentals of Quantum Field Theory (QFT), first taking a section to introduce the subject of Classical Field Theory, upon which it builds, and then providing a brief but fairly comprehensive description of the process of canonical quantization for continuous systems. There, we take the chance to explore some nontrivial effects of vacuum energy that already appear in flat space in the paradigmatic example of the Casimir Effect, and show a first example of renormalization in this simpler context. We also present some formal aspects and apparatus of the theory that will be useful on later discussions. In Chapter 3, after reviewing some general features of curved spacetime and General Relativity, we generalize the procedures presented in the previous chapter to curved, globally hyperbolic spacetimes, laying the basic formulation of QFTCS. Thereupon, we make a thorough discussion of some of its basic features, particularly analyzing the concepts of vacuum and particles, and exploring some novel aspects in the theory's phenomenology, with special attention to the creation of particles in expanding FLRW spacetimes. We then present the notion of adiabatic vacuum, which arises in the analysis of the limits of an infinitely slow expansion, for which particle creation is suppressed and one may obtain a physically meaningful (approximate) notion of vacuum state, analogous to that in Minkowski spacetime. Chapter 4 is concerned with the more technical and convoluted problem of renormalization, which is essential to make physical sense of divergent quantities of the theory, and allow for a number of physically meaningful predictions; particularly it is necessary to obtain finite expectation values for the vacuum energy in curved spacetimes, and thus to analyze its potential effects on the dynamics of the universe. In Chapter 5, we dwell in the subject of cosmology. First, we present some basic features of standard cosmology, its observational successes, and its fundamental issues, both showing the scientific motivation and constructing the necessary framework to introduce and discuss inflation. Next, we show how field theory can account for a finite primordial inflationary period, and comment briefly on symmetry breaking

and ultimately reflects our ignorance both regarding the very early universe and what a quantum theory of gravity and matter should be like. Nonetheless, there are reasonable and *self-consistent* arguments that inflation should indeed last up until $\sim 10^{-33}s$, quite far from Planckian regimes. If one wishes to go beyond a semiclassical approach, he/she is forced to struggle with the far more intricate problem of finding an adequate and computable theory of quantum gravity; a thorough and up-to-date account of efforts in this sense can be found in (15) (particularly in section 1.3.3, regarding cosmological observables) and references therein.

and the roles it could play in the early universe. Finally, we then present basic aspects of inflationary cosmology and how it addresses the problems of standard cosmology; we concretely illustrate some quantitative features of inflation in a simplified model, within the so-called chaotic inflation scenario (4), showing how its average dynamics can give rise to an exponentially expanding phase, and how its quantized fluctuations can give rise to a (nearly) scale-invariant power-spectrum, as we observe in the CMB today. Finally, in chapter 6, we summarize a few conclusions of this work, and make our final remarks regarding the perspectives on this fascinating subject.

Also, for completeness and to keep this text as self-contained as possible, we summarize a few relevant results in the subject of distributions in Appendix A, and develop some geometrical derivations for curved spaces in Appendix B.

2 FUNDAMENTALS OF QFT IN MINKOWSKI SPACETIME

Throughout this work, we shall be primarily concerned with quantum field theory in curved spacetimes (QFTCS) and some of its cosmological consequences. In many aspects, this theory arises as straightforward generalization of the more well stablished quantum field theory (QFT) in Minkowski spacetime. Thus, we find it constructive and pedagogical to introduce many of the concepts and techniques in this simpler and more familiar framework before diving in QFTCS.

We begin in section 2.1 with a brief outline of classical field theory in Minkowski spacetime, both sketching its similarities with discrete particle mechanics – which will later ease the description of canonical quantization, drawing analogies with these simpler systems –, and introducing a few tools required for handling continuous systems, with special emphasis in distributions and functional derivatives.

In section 2.2 we review the procedure of Canonical Quantization for particle systems, and directly generalize it to field theories. The latter is then exemplified in the paradigmatic example of a real scalar field, where we explore the decomposition field modes and deduce the pivotal commutation relations for the creation and ahnilation modes $a_{\mathbf{k}}$, $a^{\dagger}_{\mathbf{k}}$. Further, we write energy-momentum observables in terms of these modes and show how vacuum energy already presents divergences in Minkowski space.

Then, in section 2.3, we explore nontrivial vacuum effects on flat space by means of the paradigmatic Casimir Effect, carrying simpler procedures of regularization and subtraction to obtain a finite, renormalized vacuum energy; we then interpret our results physically.

Finally, in sections 2.4 and 2.5, we go over some technical details in the expansions of field solutions in normal modes and, in the light of these results, formally construct and interpret many elementary Green Functions of our theory and draw their connection to vacuum expectation values of two-point functions. Both sections rely heavily on the technical apparatus of Appendix A and, although they are not essential for most direct calculation in chapter 3, they should render the subject of QFT more conceptual clarity, and operationally help with more intricate calculations in chapters 3 and 4.

2.1 Classical Field Theory

2.1.1 From Particle Systems to Relativistic Fields

In ordinary particle mechanics, one is able to derive the dynamical behaviour of a system with generalized position coordinates q_i $(i = 1, 2 \ldots N)$ and velocities $\dot{q}_i$ through a Lagrangian function $L(q, \dot{q}, t)$, by the principle of stationary action. The action functional is defined by:

$$S = S[q(t)] \equiv \int_{t_1}^{t_2} L(q, \dot{q}, t)dt. \tag{2.1}$$

By demanding that the physical trajectory $q(t)$ of the system between two arbitrary endpoints $q(t_1) = q_1$ and $q(t_2) = q_2$ is that which lends S stationary, we get the Euler-Lagrange equations of motion:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0. \tag{2.2}$$

And, by solving these equations (a set of N coupled ODE's) with a known set of initial positions and velocities, $q(t_0)$ and $\dot{q}(t_0)$, one is able to predict the complete physical information of the system, given by the trajectory $q(t)$ in configuration space.

Alternatively, in the Hamiltonian formulation, one may eliminate the dependence on the velocities in favor of their canonically conjugated momenta[A] $p_i \equiv \frac{\partial L}{\partial \dot{q}_i}$, by means of a Legendre transformation. The Hamiltonian is then defined as a function of all positions q_i and momenta p_i – in the ($2N$-dimensional) domain that is collectively known as phase space – by:

$$H(q, p, t) = \sum_{i}^{N} p_i \dot{q}_i - L(q, \dot{q}, t). \tag{2.3}$$

Making use of this definition and (2.2), one may easily derive the Hamilton equations of motion:

$$\dot{q} = \frac{\partial H}{\partial q}, \qquad \dot{p} = -\frac{\partial H}{\partial p}. \tag{2.4}$$

These are generally equivalent to (2.2) (making for a system of $2N$ *first order* ODEs, rather than N *second order* ones), although they may at times be easier to solve than the former (or vice-versa). But, much more than that, the Hamiltonian formulation

[A] In the scope of the present work, for reasons of clarity and brevity, we shall not develop further on the subtleties and complications involved when there are primary constraints, that is, when one or more of the p_i are identically null and one cannot solve for all $\dot{q}_i$ in function of p_i. This leaves out important aspects of the extremely important class of gauge fields; for the reader interested in the suitable extensions to that class, we recommend (1,2,6,7), ranging through a treatment in Classical Field Theory, Quantum Field Theory, and Quantum Field Theory in Curved Spaces.

provides us with a number of geometrical aspects in phase space, which, properly exploited, not only give an entirely new perspective on classical mechanics, but also are at the roots for the procedure of canonical quantization. In particular, one explores the canonical antissimetric bilinear form: Poisson Brackets. These are defined for a pair of functions f, g in phase space as:

$$\{f, g\} = \sum_i \frac{\partial f}{\partial q_i}\frac{\partial g}{\partial p_i} - \frac{\partial f}{\partial p_i}\frac{\partial g}{\partial q_i}. \tag{2.5}$$

Particularly, one may build these brackets for any two functions from basic building blocks, the Canonical Poisson brackets of positions q_i and momenta p_i:

$$\{q_i, q_j\} = 0 = \{p_i, p_j\}, \qquad \{q_i, p_j\} = \delta_{ij}. \tag{2.6}$$

Among other things, the Poison brackets allow for yet another form to write the time evolution of any dynamical observables in our theory, through the time translation generator, the Hamiltonian[B]. Since observables can be written as functions of phase space (plus an eventual explicit time dependance), $f(q, p; t)$, one finds that its total time derivative takes the form:

$$\frac{df}{dt} = \{f, H\} + \frac{\partial f}{\partial t}, \tag{2.7}$$

as a direct consequence of (2.4). Particularly, one may even reexpress (2.4) as:

$$\dot{q} = \{q, H\}, \qquad \dot{p} = \{p, H\}. \tag{2.8}$$

Having such results stablished for discrete particle systems, we are ultimately interested in showing how they can be extended for continuous ones[C]. The simplest possible case is that of a single scalar field: we describe its configurations at a given time t by field amplitudes at all space points, rather than by a finite number of coordinates; one may think of this as passing from a discrete label i to a continuous one $\mathbf{x}$: $q_i(t) \to \phi(\mathbf{x}, t)$ (here we have in mind fields defined in ordinary 3-dimensional flat space: $\mathbf{x} \in \mathbb{R}^3$). Thus, in the absence of any internal constraints, such systems have an infinite number of degrees of

[B] One may similarly define spatial translations from momenta, rotations from angular momenta, etc.

[C] Our primary interest, of course, is studying fields, but this analysis is equally applicable to fluid mechanics, or other continuous systems.

freedom, which may be roughly regarded as "one degree of freedom per point in space". Of course, one may also have a theory with multiple or composite fields – such that several discrete indices might be required to denote different fields or field components – having instead "N degrees of freedom per point in space". We denote discrete indices collectively by the subscript a: $\phi_a(x)$.

Now, just as in the case of particle mechanics, we want to derive the dynamical behaviour for these fields through an action principle. Analogously to equation (2.1), we define the action functional of a trajectory[D]:

$$S[\phi_a] = \int_{t_1}^{t_2} L[\phi_a, \dot{\phi}_a; t] dt, \tag{2.9}$$

whereas the Lagrangian that appears in (2.9) is no longer an ordinary function of a finite number of variables, but rather a functional of the field configurations (and its first time derivatives) at time t, which prevents us from obtaining dynamical equations straighforwardly as in (2.2), through mere partial derivatives in each degree of freedom.

Apart from that, we have so far said very little about how this Lagrangian functional may depend on field variables. In particle mechanics, we have kinetical terms, usually quadratic in velocities $\dot{q}$, and mutual interaction terms, which usually depend on the separation between particles $V(\mathbf{x_i} - \mathbf{x_j})$. The first can be quite obviously transposed to field time derivatives $\dot{q}$; as of the second, we would like to make an analogue for *local* field interactions, allowing for a dependence in field amplitude *variations*, $\phi((\mathbf{x} + \delta\mathbf{x}) - \mathbf{x}) \rightarrow \delta\mathbf{x} \cdot \boldsymbol{\nabla}\phi \propto \boldsymbol{\nabla}\phi$. Since both terms involve space-time derivatives of ϕ, they are often collectively referred to as "kinetical" in field theory. Besides these terms, one often uses one-particle potentials $U(q)$, directly dependent on the coordinates[E]; analogously, we allow here for potential terms directly proportional to field amplitudes ϕ: $V(\phi)$. With that in mind, we introduce the Lagrangian density $\mathscr{L}$ containing any of these contributions; $\mathscr{L}$ will then be a local[F] function of field amplitudes and its space-time derivatives, in terms of which we write a *spatially global Lagrangian functional* L, carrying only a time dependence:

[D] Although we are no longer speaking of position variables, we still refer to the evolution of configurations of a field in a continous time interval as its 'trajectory' or 'path'.

[E] These are often due to the effects of agents considered external to our system, such as external electrostatic potential on charges, or even a spring on a simple mechanical harmonic oscillator, whose only 'internal' dynamical variable is the particle's position. Curiously the analogous terms in field theory, most notably, mass terms $\propto \phi^2$ can often be found to emerge from the system field's interactions with 'external' fields. We will discuss this point a little further in Chapter 5.

[F] At this point, we mean 'local' not (necessarily) in the relativistic sense, but rather in the sense that the values of $\mathscr{L}$ at a given point only depends on the values of ϕ in an arbitrarily small vicinity of that point.

$$L_t[\phi_a, \dot{\phi}_a] = \int_t d^3\mathbf{x}\, \mathscr{L}(\phi_a, \partial\phi_a; \mathbf{x}, t), \tag{2.10}$$

where $\partial\phi_a$ denotes collectively spatial and temporal first derivatives. Written in terms of $\mathscr{L}$, the action looks much similar to (2.1):

$$S[\phi_a] = \int_{t_1}^{t_2} dt \int_t d^3\mathbf{x}\, \mathscr{L}(\phi_a, \partial\phi_a; \mathbf{x}, t), \tag{2.11}$$

as there are now only a finite number of variables associated to each space and time points (those being the field amplitudes, as well as a finite number of derivatives, at that point) and an eventual explicit dependency on space and/or time. Thus, one may think of (2.11) as merely a 'version with extra integration dimensions' of (2.1). Thereby, it can be extremized in a similar fashion, yielding:

$$\frac{\partial}{\partial}\left(\frac{\partial\mathscr{L}}{\partial\dot{\phi}_a}\right) + \boldsymbol{\nabla}\cdot\left(\frac{\partial\mathscr{L}}{\partial(\boldsymbol{\nabla}\phi_a)}\right) - \frac{\partial\mathscr{L}}{\partial\phi_a} = 0. \tag{2.12}$$

By comparing this equation to (2.2), we see that, whereas the former is a second-order ODE system, this is a second-order PDE system. Besides being more technically complicated, these equations require not only initial conditions at some time t_0, $(\phi(t_0), \dot{\phi}(t_0))$, but ofen also spatial boundary conditions that restrain the physically permitted configurations of the theory. We shall discuss these further in the concrete example of a scalar field.

Of course, we shall be mainly interested in relativistically covariant theories. This imposition is actually simple to implement in an extreme-action (Lagrangian) formulation. All that we must require is that S is invariant under any spacetime transformations (*i.e.*, translations, rotations and boosts), that is, we must require that S is a scalar (in the relativistic sense). Since our theory is also required to be local, this means that $\mathscr{L}$ must be a *scalar field*[G]. By looking at the form of (2.11), we see that we are already integrating in spacetime, so now we express spacetime events by a single variable $x = (\mathbf{x}, t)$ and write covariantly:

$$S[\phi_a] = \int d^4x \mathscr{L}(\phi_a, \partial\phi_a; x), \tag{2.13}$$

which is extremized in an identical manner to (2.11), yielding, in Einstein summation convention:

[G] Depending on how it is defined, it could be required to be a *scalar density*; we make a more careful discussion of that point in the next chapter. For now, we are only making use of global inertial coordinates in Minkowski spacetimes, so that the two will coincide.

$$\partial_\mu\left(\frac{\partial\mathscr{L}}{\partial(\partial_\mu\phi_a)}\right)-\frac{\partial\mathscr{L}}{\partial\phi_a}=0. \tag{2.14}$$

2.1.2 Functional Derivatives

From the beginning of this chapter, we have been working with functionals, but we have only scratched very superficially what they are, and how to operate with them; so far, we have just pointed at a few classical results for extremizing them (on which we have not even elaborated much, relying on the reader's familiarity with those results from analytical mechanics). In order to better exploit them, and allow for a more systematic approach to field theory, it is worth pausing here to lay down a few basic definitions, and develop some tools for operating with functionals. A much more thorough treatment of this topic, in which the present exposition is based, can be found in the final chapter of (7).

Generally speaking, a functional F is a function of functions into numbers. That is, it takes as an input a function $f \in \mathcal{F}$, $\mathcal{F}$ being an appropriate function space, and gives a numerical output associated to it (usually a real or complex number):

$$\begin{aligned} F : \mathcal{F} &\to \mathbb{R}, \mathbb{C} \\ f &\to F[f]. \end{aligned} \tag{2.15}$$

Thus, the domain of a functional is a set of functions $\mathcal{F}$, defined in their own domain and counterdomain; in the case of fields, we are generally interested in (sufficiently smooth) functions of spacetime into a finite-dimensional space (usually numbers, tensors or spinors, depending whether we have scalar, tensor or spinor fields). The examples we have used so far is the action $S[\phi_a]$, which is a scalar function of the field values in a 4-dimensional region of spacetime and the Lagrangian $L_t[\phi_a, \dot{\phi}_a]$ which is a scalar (though not it the relativistic-invariant sense) function of field values in an (equal-time) 3-dimensional surface and their first derivative in the direction orthogonal to that surface.

Operationally, it is very important to be able to evaluate variations of functionals when we vary their arguments (and particularly, to extremize them and find stationary points). For that, as with ordinary functions with a finite-dimensional domain, one must be able to take derivatives. However, there are complications in extending this procedure to an infinite-dimensional domain, acutely so in the continuum, which do not allow for a straightforward application of mere partial derivatives. A proper extension of the concept of derivatives into functional spaces is given by the so-called *functional derivatives.* To define them, we start from the concept of directional derivatives in a finite-dimension domain:

$$\lim_{\epsilon \to 0} \frac{f(\mathbf{x} + \epsilon \mathbf{u}) - f(\mathbf{x})}{\epsilon} = \mathbf{u} \cdot \nabla f(\mathbf{x}) = \sum_i u^i \frac{\partial f}{\partial x^i}. \tag{2.16}$$

Here, we can see that the derivative ∇f is a *dual vector* that, acting on a vector $\mathbf{u}$ produces the rate of variation of f along $\mathbf{u}$ (in terms of infinitesimal variations, one may say ∇f acts on an infinitesimal displacement $\delta \mathbf{u}$ to produce the infinitesimal variation $\delta f = \delta \mathbf{u} \cdot \nabla f$).

Thus, inspired in equation (2.16), we define functional derivatives by the equation:

$$\lim_{\epsilon \to 0} \frac{F[f + \epsilon \sigma] - F[f]}{\epsilon} \equiv \int_\Omega dx \, \sigma(x) \frac{\delta F}{\delta f(x)}. \tag{2.17}$$

This produces the variations of F along (the abstract direction of) a function σ. One can also write the infinitesimal version of a variation:

$$\delta F = \int_\Omega dx \frac{\delta F}{\delta f(x)} \delta \sigma(x). \tag{2.18}$$

From the definition (2.17), one may also easily check that functional differentiation obeys some elementary identities crucial to derivative operators, such as linearity and Leibniz rule. As it happens with ordinary finite-dimensional derivatives, functional derivatives belong to the dual space $\mathcal{F}^*$ – i.e. they are *linear functionals* acting on $\mathcal{F}$ to produce numbers (namely, rates of variations along them). Thus, in general, they are *distributions*, rather than functions (see appendix A), although often one can identify them with functions.

Functionals may also depend on one or several parameters. An example above is the Lagrangian L_t, which depends on time. These are not usually treated as arguments[H], since their variations can be more straightforwardly analyzed through ordinary calculus techniques. Then, quite naturally, one may think of the function itself (evaluated at a given point) as a functional: $F_x[\phi] = \phi(x)$. This particular functional relates to a very special distribution – the Dirac delta:

$$f(x) = \int_\Omega dy f(y) \delta(x - y). \tag{2.19}$$

Comparing with our definition (2.17), we then immediately obtain that:

[H] Although, technically speaking, they are, as the functionals turn out as functions of the form $F : \mathcal{F} \times \mathbb{R}^n \to \mathbb{R}$ (where, for definiteness, we are representing real parameters and outputs).

$$\frac{\delta f(y)}{\delta f(x)} = \delta(x-y). \tag{2.20}$$

It is also worth to lay here a few operational considerations and elementary examples (for more of them, see (7)):

1- Linear functionals with an integration Kernel. They are extremely straightforward to evaluate, and one can apply (2.17) directly:

$$F_x[f] = \int_\Omega dy\, K(x,y) f(y) \quad \Rightarrow \quad \frac{\delta F_x}{\delta f(y)} = K(x,y). \tag{2.21}$$

Note also that (2.20) is just a particular application of it.

2- Locally composite functionals. These can easily be verified to obey a simple chain rule upon functional differentiation:

$$F[f] = \int_\Omega dx\, g(f(x)) \quad \Rightarrow \quad \frac{\delta F}{\delta f(x)} = \frac{dg}{df}(x); \tag{2.22}$$

3- Locally composite functionals involving a finite number derivatives. Assuming the argument functions to vanish at the boundary of the integration domain $\partial\Omega$ (or constraining the variations to always be null at $\partial\Omega$, as we do with the action), one can compute the functional derivatives through a sequence of derivations by parts, yielding:

$$\begin{aligned} F[f] &= \int_\Omega dx\, g\Big(f(x), f'(x), f''(x), ... f^{(n)}(x)\Big) \\ &\Rightarrow \quad \frac{\delta F}{\delta f(x)} = \frac{dg}{df}(x) - \frac{d}{dx}\bigg(\frac{dg}{df'}(x)\bigg) + \frac{d^2}{dx^2}\bigg(\frac{dg}{df''}(x)\bigg) ... + (-1)^n \frac{d^n}{dx^n}\bigg(\frac{dg}{df^{(n)}}(x)\bigg), \end{aligned} \tag{2.23}$$

which was written in the form of ordinary derivatives, but the extension to a finite-dimensional domain Ω is done in the obvious way in terms of partial derivarives.

Particularly, we can immediately apply this last results to write the Euler-Lagrange equations in an elegant and compact manner:

$$\frac{\delta S}{\delta \phi_a(x)} = \partial_\mu\bigg(\frac{\partial \mathscr{L}}{\partial(\partial_\mu \phi_a)}\bigg) - \frac{\partial \mathscr{L}}{\partial \phi_a} = 0. \tag{2.24}$$

2.1.3 The Hamiltonian Formulation of Field Theory

Although the Lagrangian formalism suffices for us to obtain the dynamical equations for the field in a simple and manifestly covariant manner, it is quite more complicated to obtain a quantized theory from it. Although in section 4.3 we will make a brief introduction to the Lagrangian-based path integral approach to quantum mechanics, it proves most convenient in a first approach to introduce a Hamiltonian formalism, which allows for the construction of the more straightforward scheme of canonical quantization, similarly to how it is usually done in ordinary quantum mechanics.

A disadvantage in the Hamiltonian formulation is that, unlike its Lagrangian counterpart, it must 'break' manifest spacetime covariance: in order to extract field instantaneous configurations and velocities (or momenta) from its spacetime trajectory, one must single out one time coordinate t and set it apart from space coordinates $\mathbf{x}$. For a given choice for the split of space and time, we define the field's velocities $\dot{\phi}_a(x) \equiv \frac{d}{dt}\phi_a(x)$ and momenta:

$$\pi^a(x) \equiv \frac{\partial \mathscr{L}}{\partial \dot{\phi}_a(x)}. \tag{2.25}$$

Once again, performing a Legendre transformation, we define the Hamiltonian density in phase space:

$$\mathcal{H}(\phi_a, \pi^a, x) = \pi^a(x)\phi_a(x) - \mathscr{L}(\phi_a, \dot{\phi}_a, x), \tag{2.26}$$

where the velocities $\dot{\phi}_a$ are implied to be a function of the momenta π^a.

In direct analogy with (2.5), we also define Poisson Brackets in the continuum:

$$\{F, G\} = \int d^3\mathbf{x}\, \frac{\delta F}{\delta \phi_a(\mathbf{x}, t)} \frac{\delta G}{\delta \pi^a(\mathbf{x}, t)} - \frac{\delta G}{\delta \phi_a(\mathbf{x}, t)} \frac{\delta F}{\delta \pi^a(\mathbf{x}, t)}\ . \tag{2.27}$$

Particularly, we have the fundamental canonical Poisson Brackets:

$$\{\phi_a(\mathbf{x}, t), \pi^b(\mathbf{y}, t)\} = \delta_a{}^b\, \delta^{(3)}(\mathbf{x} - \mathbf{y}), \tag{2.28}$$

which will play a key role in canonical quantization.

2.1.4 The free real scalar field

An example of major importance which we shall explore in extensive detail throughout this work is the free real scalar field (also known as the real Klein-Gordon field). The starting point for defining it is the Lagrangian density:

$$\mathscr{L} = \frac{1}{2}\eta^{\mu\nu}(\partial_\mu\phi)(\partial_\nu\phi) - \frac{m^2}{2}\phi^2. \tag{2.29}$$

This contains ordinary kinetic terms as well as a simple quadratic (harmonic) potential term. Here, m^2 is a positive parameter characterizing the steepness of the potential well (we use this suggestive notation because, as we shall see later, m will be identified with the mass of the quanta – the particles – of the quantized field); we also note that, in natural units ($\hbar{=}c{=}1$), m has units of inverse length[I], conferring the field with a characteristic lengh scale, m^{-1}.

From this Lagrangian, we easily obtain the dynamical equations:

$$[\Box + m^2]\phi = 0, \tag{2.30}$$

where we have defined the D'Alembertian in flat spacetime: $\Box\phi = \eta^{\mu\nu}\partial_\mu\partial_\nu\phi = \partial_t^2\phi - \nabla^2\phi$ (where the last equality is expressed in globally inertial Cartesian coordinates, and ∇^2 represents an ordinary spatial Laplacian: $\nabla^2 = \partial_x^2 + \partial_y^2 + \partial_z^2$).

Since the field equations are linear, we may expand any solutions in terms of a complete set of modes. One particularly convenient basis for (2.30) are plane wave modes:

$$u_{\mathbf{k}}(\mathbf{x}, t) = \frac{1}{\sqrt{2\omega_k}} e^{-i\omega_k t} e^{i\mathbf{k}\,\cdot\,\mathbf{x}}. \tag{2.31}$$

Here, we have defined positive frequencies, $\omega_k = +\sqrt{m^2 + \mathbf{k}^2}$, and we have included the normalization factor $(2\omega_k)^{-1/2}$ for later convenience. We may then write the field expansion as:

$$\phi(\mathbf{x}, t) = \sum_{\mathbf{k}} a_{\mathbf{k}} u_{\mathbf{k}}(\mathbf{x}, t) + a_{\mathbf{k}}^* u_{\mathbf{k}}^*(\mathbf{x}, t). \tag{2.32}$$

Of course, to specify which range of wave vectors $\mathbf{k}$ are allowed (or, more generally, which combinations of $a_{\mathbf{k}}$ and $a_{\mathbf{k}}^*$ are permitted), we must also specify (spatial) boundary conditions. These should, of course, depend on the global physical conditions we want to impose to our field. Whereas these are relatively straightforward for spatially bounded systems (a paradigmatic example in field theory is the electromagnetic field confined within a conducting surface, for which one just applies Dirichlet conditions on the boundary), they may raise nontrivial questions for spatially open systems (as will be the case in

[I] In regular units, this inverse lenght reads $\mu = \frac{mc}{\hbar}$.

cosmological contexts with noncompact universes), regarding the behaviour of the field at infinity. Nevertheless, such questions are often of little relevance to the local dynamics[J], so it is a common practice to take artificial boundary conditions that simplify one's calculations. A rather convenient choice is to take periodic boundary conditions in a cube with dimentions $L \times L \times L$, such that the wave vectors $\mathbf{k}$ (and therefore the frequencies ω_k) only take a discrete set of values; in a properly chosen Cartesian grid, their components will be:

$$k^i = \frac{2\pi}{L} n_i, \qquad i = 1, 2, 3\,, \qquad n_i \in \mathbb{Z}. \tag{2.33}$$

In doing so, we also incorporate a volume factor $V = L^3$ in the normalization of (2.31), defining:

$$u_{\mathbf{k}}(\mathbf{x}, t) \equiv \frac{1}{\sqrt{2V\omega_k}} e^{-i\omega_k t} e^{i\mathbf{k}\,\cdot\,\mathbf{x}}. \tag{2.34}$$

From this construction, a very straightforward way to analyze the continuum limit and drop the periodic conditions is to take $L \to \infty$, and make a proper change in normalization (this amounts simply to going from a discrete Fourier series to a continuous Fourier transform; see e.g. (20)). The adjustment that we make in the latter aims at bilinear integrals (particularly the orthornormality conditions (2.43) (2.45) below) and can be motivated as follows: for finite L, one has the spectral volume around each mode (the "volumetric spacing between modes"): $\Delta^3\mathbf{k} = (2\pi/L)^3$. Thus, we take sums into integrals by making:

$$\left(\frac{2\pi}{L}\right)^3 \sum_{\mathbf{k}} \;=\; \sum_{\mathbf{k}} \Delta^3\mathbf{k} \longrightarrow \int d^3\mathbf{k}\,. \tag{2.35}$$

We then end up with the following normalized modes in the continuum:

$$u_{\mathbf{k}}(\mathbf{x}, t) \equiv \frac{1}{\sqrt{2(2\pi)^3\omega_k}} e^{-i\omega_k t} e^{i\mathbf{k}\,\cdot\,\mathbf{x}}, \tag{2.36}$$

for which one writes the integral field expansion:

$$\phi(\mathbf{x}, t) = \int d^3\mathbf{k}\, a(\mathbf{k}) u_{\mathbf{k}}(\mathbf{x}, t) + a^*(\mathbf{k}) u^*_{\mathbf{k}}(\mathbf{x}, t). \tag{2.37}$$

[J] One should bear in mind, however, that quantum theory has important nonlocal features. We shall further analyze the effects of boundary conditions when we discuss the Casimir Effect in section 2.3.

Note this particular approach to the continuum also implies a boundary condition at infinity: by restricting the wave vectors to be real, it forces $u_{\mathbf{k}}$ to remain bounded, not allowing for any exponentially increasing solutions. This will be crucial for mode decomposition, as we want to constrain our modes to have finite projections on integrable field solutions (*e.g.*, wave-packets).

In either case, to obtain the complete physical information about this system, one must solve the field equations, with some given initial conditions[K]. Well, given the expansion (2.32) (or (2.37)), this amounts to finding the coefficients $a_{\mathbf{k}}$ – i.e. *the amplitudes for each field mode* (which do not change with time, since the modes are decoupled and evolve independently) – for which we match the initial conditions:

$$\phi(\mathbf{x}, t_0) = f(\mathbf{x}), \qquad \dot{\phi}(\mathbf{x}, t_0) = g(\mathbf{x}). \tag{2.38}$$

Note that the (spectral) mode amplitudes $a_{\mathbf{k}}$ and the (spatial) field amplitudes $\phi(\mathbf{x}, t)$ depend linearly on one another. Since we can already express $\phi(\mathbf{x}, t)$ in terms of $a_{\mathbf{k}}$, the above task amounts to inverting that expression to obtain $a_{\mathbf{k}}$ in terms of $\phi(\mathbf{x}, t)$ (and $\dot{\phi}(\mathbf{x}, t)$) at $t = t_0$. More specifically, we want to *project* the initial conditions in all the basis modes. To achieve this, it proves useful to define the following scalar product (sometimes called the Klein-Gordon product):

$$(\phi, \psi) \equiv i \int_t d^3\mathbf{x}\; \phi^* \overleftrightarrow{\partial_t} \psi = i \int_t d^3\mathbf{x}\; \phi^* \partial_t \psi - (\partial_t \phi^*) \psi, \tag{2.39}$$

where the integration may be taken in any arbitrary fixed time t. One can immediately verify it obeys the following elementary properties:

$$(u, \alpha v_1 + v_2) = \alpha (u, v_1) + (u, v_2), \qquad \alpha \in \mathbb{C}, \tag{2.40a}$$

$$(v, u) = -(u^*, v^*) = (u, v)^*. \tag{2.40b}$$

Note, in particular, that (2.40b) implies:

$$(u, u^*) = 0.$$

The definition (2.39) suggests that this product is time-dependent (i.e. that it depends on the equal-time surface Σ_t chosen to perform the integration). Indeed, that

[K] Or boundary and initial conditions, but, as we just mentioned, spatial boundary conditions are usually incorporated in the determination of a complete set of modes would be the case if we calculated the product of two (completely) arbitrary functions of spacetime. However, we shall show that the product of any two *solutions of the dynamical equations* is actually conserved (presently, we limit our demonstration to equal-time surfaces Σ_t; we shall give this demonstration in greater generality in chapter 3).

Let us take 2 time instants $t' > t$. Note that the 2 hypersurfaces $\Sigma_{t'}$ and Σ_t are the boundary of the spacetime region Ω between them, and that their outward-pointing normal vectors are $n^a(\frac{\partial}{\partial t})^a$ and $n^a = -(\frac{\partial}{\partial t})^a$, respectively, so that the difference between the product (ϕ, ψ) computed at t' and t may be written as a boundary term:

$$\begin{aligned}(\phi,\psi)_{t'} - (\phi,\psi)_t &= \int_{\Sigma_{t'}} d^3\mathbf{x} n^\mu \phi^*(x) \overleftrightarrow{\partial_\mu} \psi(x) - \int_{\Sigma_t} d^3\mathbf{x} n^\mu \phi^*(x) \overleftrightarrow{\partial_\mu} \psi(x) \\ &= \int_{\partial\Omega} d^3\mathbf{x} n^\mu \phi^* \overleftrightarrow{\partial_\mu} \psi.\end{aligned} \tag{2.41}$$

Then, applying the Gauss divergence theorem, we obtain:

$$\begin{aligned}(\phi,\psi)_{t'} - (\phi,\psi)_t &= \int_\Omega d^4x \partial^\mu \Big(\phi^*(x) \overleftrightarrow{\partial_\mu} \psi(x)\Big) \\ &= \int_\Omega d^4x \Big[\phi^*(x)\Box_x\psi(x) - \psi(x)\Box_(x)\phi^*(x)\Big] \\ &= \int_\Omega m^2 \Big[\phi^*(x)\psi(x) - \psi(x)\phi^*(x)\Big] \\ &= 0.\end{aligned} \tag{2.42}$$

With these basic properties in mind, we may now use this scalar product to solve the initial value problem of the free scalar field. First, note that the field modes (2.34) are orthornormal with respect to it:

$$(u_\mathbf{k}, u_{\mathbf{k}'}) = \delta_{\mathbf{k}\mathbf{k}'} = -(u^*_\mathbf{k}, u^*_{\mathbf{k}'}), \tag{2.43a}$$

$$(u_\mathbf{k}, u^*_{\mathbf{k}'}) = 0, \tag{2.43b}$$

so that the coefficients of expansion (2.32) are simply given by the projections in each field mode:

$$a_\mathbf{k} = (u_\mathbf{k}, \phi), \tag{2.44a}$$

$$a^*_\mathbf{k} = -(u^*_\mathbf{k}, \phi), \tag{2.44b}$$

which can be directly computed at the time t_0, using the inicial conditions (2.38).

In the continuum case, one can similarly verify that the modes also follow a suitable orthornormality condition:

$$(u(\mathbf{k}), u(\mathbf{k}')) = \delta(\mathbf{k} - \mathbf{k}') = -(u^*(\mathbf{k}), u^*(\mathbf{k}')), \tag{2.45a}$$

$$(u(\mathbf{k}), u^*(\mathbf{k}')) = 0, \tag{2.45b}$$

and the coefficents can be similarly obtained by the projections:

$$a(\mathbf{k}) = (u(\mathbf{k}), \phi), \tag{2.46a}$$

$$a^*(\mathbf{k}) = -(u^*(\mathbf{k}), \phi). \tag{2.46b}$$

Presently, we limit our presentation of the field modes to just these basic properties. We shall explore them in greater detail for the quantized field, where they will play a central role in our Fock Space representation.

2.2 Canonical Quantization

Now that we developed some key aspects of the formalism for classical fields, we would like to proceed to their quantization. Here, we shall carry this procedure in the Hamiltonian formalism, drawing close analogy to discrete particle systems. Recall that in canonical quantization of a particle system with position coordinates x_i and canonically cojugated momenta p_i we promote these classical observables to quantum *operators* – linear operators acting in a suitable Hilbert space $\mathcal{H}$ (usually taken to be an appropriate subspace of square-integrable functions $\mathcal{F} \subset \mathcal{L}^2(\mathbb{R}^N)$) –, obeying the canonical commutation relations ($\hbar = 1$):

$$[x_i, x_j] = [p_i, p_j] = 0, \qquad [x_i, p_j] = i\delta_{ij}. \tag{2.47}$$

These relations are postulated in direct reference to the canonical Poisson brackets (2.6), where one substitutes the brackets between two classical observables by $-i$ times the commutator between their corresponding quantum operators. At this point, we stress that these operators do not have direct physical significance on their own; truly observable quantities arise, for example, in the form of expectation values in specified quantum states $|\psi\rangle$, such as $\langle\psi|x_i|\psi\rangle$, as well as of projections between two states in a given time $\langle\phi|\psi\rangle$ (which yield transition amplitudes between the states $|\psi\rangle$ and $|\phi\rangle$); more generally, one can consider transition amplitudes of the form $\langle\phi|A|\psi\rangle$.

Thus, there is an inherent ambiguity in this formalism concerning the definition of state vectors and operators. By performing complementary unitary transformations on both, transforming all state vectors as $|\psi\rangle \to U|\psi\rangle$ (and dual vectors as $\langle\psi| \to \langle\psi| U^\dagger$) and operators as $A \to UAU^\dagger$ ($UU^\dagger = \mathbb{1}$), one attains an equivalent physical description of the theory, as all observable quantities remain invariant. Each of these descriptions corresponds to a representation, or *picture* of the theory.

Two eminent pictures of quantum theory are the Schrödinger and the Heisenberg pictures. In nonrelativistic quantum mechanics, one usually works in the Schrödinger picture, in which the 'fundamental' observables x_i, p_i are time-independent and state vectors evolve according to the Schrödinger equation $i\frac{d}{dt}|\psi(t)\rangle = H|\psi(t)\rangle$, whose solution for a given initial state $|\psi_0\rangle$ at $t = 0$ is given in terms of the unitary evolution operator $U(t)$:

$$|\psi(t)\rangle = U(t)|\psi_0\rangle\,, \tag{2.48}$$

where U obeys the following relations:

$$U^\dagger U = UU^\dagger = \mathbb{1}, \qquad i\frac{d}{dt}U(t) = HU(t), \qquad U(0) = \mathbb{1}. \tag{2.49}$$

Particularly, in the case where H is time-independent, we recover the simple form: $U(t) = e^{-iHt}$.

On the other hand, in the Heisenberg picture, state vectors are kept fixed and the operators evolve in time, in such a manner that all physical observables remain unchanged:

$$|\psi^H\rangle = U^\dagger(t)|\psi^S(t)\rangle = |\psi_0\rangle\,, \qquad x^H(t) = U^\dagger(t)x^S U(t). \tag{2.50}$$

Taking the time derivative of the Heisenberg observables, we see that they obey formally identical equations to their classic counterparts (2.8) (but which must now be interpreted as operator-valued equations):

$$i\frac{d}{dt}x(t) = [x(t), H], \qquad\qquad i\frac{d}{dt}p(t) = [p(t), H]. \tag{2.51}$$

As in the classical context, these will be ultimately equivalent to the Euler-Lagrange (operator-valued) equations:

$$\frac{d}{dt}\left(\frac{dL}{d\dot{x}}\right) - \frac{dL}{dx} = 0. \tag{2.52}$$

Finally, by re-evaluating the commutation relations (2.47), one may easily verify that they still hold for the Heisenberg operators *for equal times*:

$$[x_i(t), x_j(t)] = [p_i(t), p_j(t)] = 0, \qquad [x_i(t), p_j(t)] = i\delta_{ij}. \tag{2.53}$$

We may then take a similar approach for quantizing field systems. We promote the classical field ϕ to a quantum *field operator*, for which – based on the classical Poisson brackets (2.28)– we postulate the so called *equal-time commutation relations*:

$$[\phi_a(\mathbf{x}, t), \phi_b(\mathbf{y}, t)] = [\pi^a(\mathbf{x}, t), \pi^b(\mathbf{y}, t)] = 0, \qquad [\phi_a(\mathbf{x}, t), \pi^b(\mathbf{y}, t)] = i\delta_a^{\ b}\, \delta^{(3)}(\mathbf{x} - \mathbf{y}). \tag{2.54}$$

These relations – although not manifestly covariant, as they require singling out a time t – will actually be Lorentz invariant, provided we have a Lorentz invariant (scalar) Lagrangian density. (For a derivation of covariant commutation relations for various fields in flat spacetime, see for example (6).) We just state the result here that, in this case, eqs. (2.54) will imply:

$$[\phi(x), \phi(y)] = 0, \tag{2.55}$$

always that x and y have a spacelike separation.

As position observables in particle mechanics, the quantized fields obey, in the Heisenberg picture, analogous (operator-valued) equations as their classical counterparts. In the latter case, however, we often only appeal to a Hamiltonian formalism to outline the bridge in canonical quantization, being more convenient to work directly with the Euler-Lagrange equations to analyze the fields' dynamics.

Finally, we remind that, while in classical mechanics one can in principle completely determine the values of positions and momenta simultaneously at a given time, and thus predict with certainty their values for any other time, this is forbidden in quantum mechanics in virtue of the uncertainty principle. In the latter case, the maximal information that one can ascertain, which can be used to completely determine a *quantum state* is attached to the so-called *Complete Sets of Commuting Observables* (or C.S.C.O.'s), whose eigenstates spawn the entire Hilbert Space $\mathcal{H}$. The most immediate such C.S.C.O's are those given *either* by positions *or* momenta (field configurations or field momenta) at any given time. We write, for instance (in the Heisenberg Picture):

$$|\phi'\rangle : \phi(\mathbf{x}, t_0)\,|\phi'\rangle = \phi'(\mathbf{x}, t_0)\,|\phi'\rangle\,, \quad \forall \mathbf{x} \in \mathbb{R}^3, \tag{2.56}$$

$$|\pi'\rangle : \pi(\mathbf{x}, t_0)\,|\pi'\rangle = \pi'(\mathbf{x}, t_0)\,|\pi'\rangle\,, \quad \forall \mathbf{x} \in \mathbb{R}^3, \tag{2.57}$$

where t_0 is *fixed* time, and we have used primes to distinguish field/momentum *eigenvalues* from field/momentum operators[L].

Of course, one can build a myriad of different C.S.C.O's in $\mathcal{H}$, usually by considering combinations and functions of positions and momenta. In the next section, we shall particularly emphasize those associated with field modes.

2.2.1 Quantizing the scalar field

Now that we are armed with a general prescription to quantize field systems we shall illustrate it explicitly in the case of the scalar field ϕ (2.29). After obtaining the quantized field, we shall give emphasis to its expansion in normal modes and construct the Fock space based on them, noting how the notion of particles naturally emerges as excitations of these field modes. We also use this expansion to compute energy-momentum observables.

As ϕ obeys the operator analogous of equations (2.30), it remains useful to expand the field in plane wave modes (2.31). However, we must substitute the classical amplitudes by the field mode operators $a_{\mathbf{k}}$ and $a^\dagger_{\mathbf{k}}$:

$$\phi(\mathbf{x}, t) = \sum_{\mathbf{k}} a_{\mathbf{k}} u_{\mathbf{k}}(\mathbf{x}, t) + a^\dagger_{\mathbf{k}} u^*_{\mathbf{k}}(\mathbf{x}, t), \tag{2.58}$$

$$\pi(\mathbf{x}, t) = \sum_{\mathbf{k}} -i\omega_k \Big(a_{\mathbf{k}} u_{\mathbf{k}}(\mathbf{x}, t) - a^\dagger_{\mathbf{k}} u^*_{\mathbf{k}}(\mathbf{x}, t)\Big). \tag{2.59}$$

Such expansions express the entire range of canonical observables (ϕ and π, $\forall \mathbf{x}$) in terms of a and $a^\dagger$ ($\forall \mathbf{k}$). Conversely, by using the projections $a_{\mathbf{k}} = (u_{\mathbf{k}}, \phi)$ – which involve both ϕ and its time derivative –, we may express a and $a^\dagger$ in terms of ϕ and π. Then we can easily derive:

$$\begin{aligned}
[a_{\mathbf{k}}, a_{\mathbf{k}'}] &= \Big[\int_t d^3\mathbf{x}(u^*_{\mathbf{k}}(x)\pi(x) - i\omega_k u^*_{\mathbf{k}}(x)\phi(x)), \int_t d^3\mathbf{y}(u^*_{\mathbf{k}'}(y)\pi(y) - i\omega_{k'} u^*_{\mathbf{k}'}(y)\phi(y))\Big] \\
&= \int_t d^3\mathbf{x} \int_t d^3\mathbf{y}\, u^*_{\mathbf{k}}(x) u^*_{\mathbf{k}'}(y) \Big\{ -i[\pi(\mathbf{x}, t), \phi(\mathbf{y}, t)] - i[\phi(\mathbf{x}, t), \pi(\mathbf{y}, t)] \\
&\qquad\qquad + [\phi(\mathbf{x}, t), \phi(\mathbf{y}, t)] + i^2[\pi(\mathbf{x}, t), \pi(\mathbf{y}, t)] \Big\}
\end{aligned}$$

[L] Rigorously speaking, $|\phi'\rangle$ and $|\pi'\rangle$ are not rigorously states belonging to $\mathcal{H}$, but they can be used to spawn any actual states $|\Psi\rangle \in \mathcal{H}$. See appendix A for more details.

$$\begin{aligned} &= \int_t d^3\mathbf{x} \int_t d^3\mathbf{y} u^*_{\mathbf{k}}(x) u^*_{\mathbf{k}'}(y) \Big\{ \delta(\mathbf{x}-\mathbf{y}) - \delta(\mathbf{x}-\mathbf{y}) \Big\} \\ &= 0, \end{aligned} \tag{2.60}$$

where we have used the time invariance of (2.39) to compute the projections for $a_{\mathbf{k}}$ and $a^\dagger_{\mathbf{k}}$ at the same time t, and be able to use (2.54).

Similarly, one can compute $[a^\dagger_{\mathbf{k}}, a^\dagger_{\mathbf{k}'}] = 0$. We then have the nontrivial commutator:

$$\begin{aligned} [a_{\mathbf{k}}, a^\dagger_{\mathbf{k}'}] &= \Big[\int_t d^3\mathbf{x}(u^*_{\mathbf{k}}(x)\pi(x) - i\omega_k u^*_{\mathbf{k}}(x)\phi(x)), -\int_t d^3\mathbf{y}(u_{\mathbf{k}'}(y)\pi(y) - i\omega_{k'} u_{\mathbf{k}'}(y)\phi(y))\Big] \\ &= \int_t d^3\mathbf{x} \int_t d^3\mathbf{y}\, u^*_{\mathbf{k}}(x) u_{\mathbf{k}'}(y) \Big(i\omega_{k'}[\pi(\mathbf{x},t), \phi(\mathbf{y},t)] - i\omega_k[\phi(\mathbf{x},t), \pi(\mathbf{y},t)] \Big) \\ &= \int d^3\mathbf{x}\, u^*_{\mathbf{k}}(x) u^*_{\mathbf{k}'}(x)(\omega_{k'} + \omega_k) \\ &= \delta_{\mathbf{k}\mathbf{k}'}. \end{aligned} \tag{2.61}$$

Summarizing:

$$[a_{\mathbf{k}}, a_{\mathbf{k}'}] = [a^\dagger_{\mathbf{k}}, a^\dagger_{\mathbf{k}'}] = 0, \qquad [a_{\mathbf{k}}, a^\dagger_{\mathbf{k}'}] = \delta_{\mathbf{k}\mathbf{k}'}. \tag{2.62}$$

Classically, a complete set of observables for this field was given by field amplitudes at all spacetime points (which, in their turn, could be obtained by the field amplitudes and their first time derivavatives *for all space points at a given time*, *i.e.*, at surface of simultaneity); equivalently, one could specify the field amplitudes $a_{\mathbf{k}}$ for all modes. For the quantized field, however, it is impossible to determine such complete information about the field's trajectory (and thus, about its amplitudes and 'velocities' at a given time); conversely, in the mode perspective, one cannot attain the full information about the amplitudes $a_{\mathbf{k}}$ (technically, $a_{\mathbf{k}}$ is not an observable in the traditional quantum mechanical sense, since it is not a self-adjoint operator, and thus there is no guarantee that it will be possible to build a basis of eigenstates in $\mathcal{H}$ from it[M]). That is, one cannot find simultaneously its real and imaginary parts (note that, for a simple harmonic oscillator, $\mathrm{Re}(a) \propto x$ and $\mathrm{Im}(a) \propto p$). Alternatively, one can decompose $a_{\mathbf{k}}$ in its magnitude $|a_{\mathbf{k}}| = (a_{\mathbf{k}} a^*_{\mathbf{k}})^{\frac{1}{2}}$ and phase $\theta_{\mathbf{k}}$, which, again, cannot be determined simultaneously. In QFT it is their magnitude that takes a proeminent role[N], more precisely, their quadratic magnitude, which classically reads $|a_{\mathbf{k}}|^2 = a_{\mathbf{k}} a^*_{\mathbf{k}}$; then, in the quantized theory, we define the occupation observables:

[M] That is not to say that there are no Eingestates of $a_{\mathbf{k}}$ in $\mathcal{H}$. In fact, the so-called coherent states, taking the form $|\nu\rangle = e^{\nu a^\dagger}|0\rangle$, not only form a continuous family of eingenstates of a, $a\,|\nu\rangle = \nu\,|\nu\rangle$, but they are also very important in the evaluation of the classical limit of the theory; For more details on them, check chapter 1 of (6) (particularly, exercise 1.1) .

[N] For a comprehensive discussion of phase observables, which seldom appear in the QFT literature, see *e.g.* (chapter 5 of) (30).

$$N_{\mathbf{k}} = a_{\mathbf{k}} a_{\mathbf{k}}^{\dagger}. \tag{2.63}$$

It follows from the commutation relations (2.62) that our modes decompose the scalar field as an infinite set of decoupled harmonic oscillators, whence we have immediately that the spectrum for each occupation observable is just natural numbers, *i.e.*, $\sigma(N_{\mathbf{k}}) = \mathbb{N}$; for this reason, these are often called occupation numbers. We can also define the *total* occupation number $N \equiv \sum_{\mathbf{k}} N_{\mathbf{k}}$, which will similarly have the spectrum $\sigma(N) = \mathbb{N}$. The description in terms of occupation numbers then gives us a natural framework to define *particles* in our theory: we interpret each quantum in mode $u_{\mathbf{k}}$ as a particle of momentum $\mathbf{k}$, and energy $E = \sqrt{m^2 + \mathbf{k}^2}$ (this also gives us a natural interpretation of m as the mass of each particle). This representation of our Hilbert Space, based on the occupation numbers for each mode, is called a *Fock Space*. We can construct it starting from the vacuum state, $|0\rangle$, which has no quanta (particles) of any type. More precisely, it is defined by:

$$a_{\mathbf{k}} |0\rangle = 0, \quad \forall \mathbf{k}. \tag{2.64}$$

Then, we can obtain the all n-particle states by applying the various creation operators $a_{\mathbf{k}}^{\dagger}$ to the vaccuum:

$$|n_{\mathbf{k_1}}, n_{\mathbf{k_2}} ...\rangle = \frac{1}{\sqrt{n_{\mathbf{k_1}}! n_{\mathbf{k_2}}! ...}} (a_{\mathbf{k_1}}^{\dagger})^{n_{\mathbf{k_1}}} (a_{\mathbf{k_2}}^{\dagger})^{n_{\mathbf{k_2}}} ... |0\rangle . \tag{2.65}$$

Furthermore, we can expand any observables of our theory in terms of field modes. Special dynamic interest attaches to energy and momentum of our field. In relativistic theories in the continuum, energy, momentum and stress are all codified in a single tensorial observable, the well-known stress tensor $T_{\mu\nu}$[O]. Its time-time and time-space components are interpreted as energy and momentum density, respectively, whereas its space-space components are related to stresses (being its diagonal components related to pressures, and its non-diagonal ones, to shears), all with respect to a stationary observer in the adopted coordinate frame. That is, given an observer $\mathcal{O}$ with a normalized vector u^a tangent to its worldline, and normalized 3-vectors e_i^a:

[O] We shall not construct $T_{\mu\nu}$ here in detail. If the reader is unfamiliar with it, we suggest chapter 10 of (7) for the construction of a classical $T_{\mu\nu}$ in nonrelativistic and special relativistic theories, as a conserved Noether current associated with the symmetries of space-time translations. See also chapter 4 of (5) for a discussion of $T_{\mu\nu}$ in Special and General Relativity.

$$\begin{cases} T_{\mu\nu}u^{\mu}u^{\nu} \equiv T_{00} = \mathcal{H} = \rho, & (2.66a) \\ T_{\mu\nu}u^{\mu}e^{\nu}_{i} \equiv T_{0i} = \pi_i, & (2.66b) \\ T_{\mu\nu}e^{\mu}_{i}e^{\nu}_{j} \equiv T_{ij}, & (2.66c) \end{cases}$$

being the diagonal spatial components, T_{ii} related to pressures p_i in each direction (for an nonviscous isotropic fluid or field, $T_{ij} = p\delta_{ij}$). We urge the reader not to mistake the *mechanical 3-momentum density* π^i with the *canonically conjugated momentum* π.

In virtue of Noether's Theorem (7), $T_{\mu\nu}$ is a conserved current, that is:

$$\partial_{\mu}T^{\mu}{}_{\nu} = 0. \tag{2.67}$$

Using Gauss's Theorem, it is straightforward to verify that (less of any energy-momentum currents at infinity), this will also imply a *global* conservation law in Minkowski spacetime:

$$P^{\mu}(t) = \int_t d^3\mathbf{x}\, n^{\mu}T^{\mu}{}_{\nu}(\mathbf{x},t) = \int_t d^3\mathbf{x}\, T^{\mu}{}_{0}(\mathbf{x},t) \qquad \Rightarrow \qquad \frac{d}{dt}P^{\mu}(t) = 0, \tag{2.68}$$

where n^{ν} is the future-directed normal to the equal-time surface Σ_t. The spatial and temporal components of P^{μ} are interpreted as the total energy and momentum:

$$\begin{cases} P^0 = \int d^3\mathbf{x}\, \mathcal{H}(\mathbf{x},t) = H, & (2.69a) \\ P^i = \int d^3\mathbf{x}\, \pi^i(\mathbf{x},t). & (2.69b) \end{cases}$$

For a free scalar field, the canonical stress tensor reads (5, 7):

$$T_{\mu\nu} = (\partial_{\mu}\phi)(\partial_{\nu}\phi) - \tfrac{1}{2}\eta_{\mu\nu}\Big[(\partial^{\alpha}\phi)(\partial_{\alpha}\phi) - m^2\phi^2\Big]. \tag{2.70}$$

Particularly, we have the energy density,

$$\mathcal{H} = \frac{1}{2}\Big(\pi^2 + (\nabla\phi)^2 + m^2\phi^2\Big), \tag{2.71}$$

and the 3-momentum density $\boldsymbol{\pi}$ (again, do not misktake it for the conjugated momentum π):

$$\boldsymbol{\pi} = \dot{\phi}\,(\boldsymbol{\nabla}\phi) = \pi\,(\boldsymbol{\nabla}\phi). \tag{2.72}$$

To evaluate their global (spatially integrated) correspondents (2.69), we note that the integral for each of their terms can be computed without much difficulty by noting that:

$$\frac{1}{V}\int d^3\mathbf{x} e^{i(\mathbf{k}-\mathbf{k}')\,\cdot\,\mathbf{x}} = \delta_{\mathbf{k},\mathbf{k}'}, \qquad \frac{1}{(2\pi)^3}\int d^3\mathbf{x} e^{i(\mathbf{k}-\mathbf{k}')\,\cdot\,\mathbf{x}} = \delta(\mathbf{k}-\mathbf{k}'), \tag{2.73}$$

in the dicrete and in the continuum, respectively. For simplicity, we work in the discrete, for which we find the basic terms of the Hamiltonian:

$$\int d^3\mathbf{x}\,\phi^2(\mathbf{x},t) = \frac{1}{2}\sum_{\mathbf{k}}\frac{1}{\omega_k}\Big(a_{\mathbf{k}}a^\dagger_{\mathbf{k}} + a^\dagger_{\mathbf{k}}a_{\mathbf{k}} + a_{\mathbf{k}}a_{-\mathbf{k}}e^{-2i\omega_k t} + a^\dagger_{\mathbf{k}}a^\dagger_{-\mathbf{k}}e^{2i\omega_k t}\Big), \tag{2.74}$$

$$\int d^3\mathbf{x}(\nabla\phi)^2(\mathbf{x},t) = \frac{1}{2}\sum_{\mathbf{k}}\frac{\mathbf{k}^2}{\omega_k}\Big(a_{\mathbf{k}}a^\dagger_{\mathbf{k}} + a^\dagger_{\mathbf{k}}a_{\mathbf{k}} + a_{\mathbf{k}}a_{-\mathbf{k}}e^{-2i\omega_k t} + a^\dagger_{\mathbf{k}}a^\dagger_{-\mathbf{k}}e^{2i\omega_k t}\Big), \tag{2.75}$$

$$\int d^3\mathbf{x}\,\dot\phi^2(\mathbf{x},t) = \frac{1}{2}\sum_{\mathbf{k}}\frac{\omega_k^2}{\omega_k}\Big(a_{\mathbf{k}}a^\dagger_{\mathbf{k}} + a^\dagger_{\mathbf{k}}a_{\mathbf{k}} + a_{\mathbf{k}}a_{-\mathbf{k}}e^{-2i\omega_k t} + a^\dagger_{\mathbf{k}}a^\dagger_{-\mathbf{k}}e^{2i\omega_k t}\Big). \tag{2.76}$$

Combining these, we obtain the total Hamiltonian:

$$H = \int d^3\mathbf{x}\,\mathcal{H}(x) = \sum_{\mathbf{k}}\omega_k(N_{\mathbf{k}} + \tfrac{1}{2}), \tag{2.77}$$

where we see that the time-dependent terms cancel out, and we have rewritten the time-independent ones in terms of $N_{\mathbf{k}}$ and the commutator $[a_{\mathbf{k}}, a^\dagger_{\mathbf{k}}] = 1$.

We can similarly compute the total 3-momentum:

$$\mathbf{P} = \int d^3\mathbf{x}\,\dot\phi(x)(\boldsymbol{\nabla}\phi) = \frac{1}{2}\sum_{\mathbf{k}}\frac{\omega_k\mathbf{k}}{\omega_k}\Big(a_{\mathbf{k}}a^\dagger_{\mathbf{k}} + a^\dagger_{\mathbf{k}}a_{\mathbf{k}} + a_{\mathbf{k}}a_{-\mathbf{k}}e^{-2i\omega_k t} + a^\dagger_{\mathbf{k}}a^\dagger_{-\mathbf{k}}e^{2i\omega_k t}\Big) \tag{2.78}$$

$$= \sum_{\mathbf{k}}\mathbf{k}\,N_{\mathbf{k}}, \tag{2.79}$$

where we have exploited the parity symmetry $\mathbf{k} \to -\mathbf{k}$ to conveniently cancel the last two terms (as well as the one emerging from the commutator) *in this conditionally convergent sum*, and write the result in the last line.

Having constructed these observables, it will be of particular interest to us to evaluate their vacuum expectation value, which should correspond to vacuum energy and momentum, respectively. Not surprisingly, we find the vacuum momentum (in our particular summation convention) to be null:

$$\langle 0|\mathbf{P}|0\rangle = \mathbf{0}. \tag{2.80}$$

However, if we attempt to evaluate the vacuum energy, we immediately find a divergent result:

$$\langle 0|H|0\rangle = \frac{1}{2}\sum_{\mathbf{k}}\omega_k, \tag{2.81}$$

as the sum extends to infinitely many modes of arbitrarily high frequencies (therefore this is called an *ultraviolet (UV) divergence*).

It is particularly useful to analyze this divergence in the continuum limit, making the correspondence (2.35), for which we obtain:

$$\langle 0|H|0\rangle = \frac{1}{2}\sum_{\mathbf{k}}\omega_k \;\longrightarrow\; \lim_{L\to\infty}\frac{L^3}{2(2\pi)^3}\int d^3\mathbf{k}\omega_k = \left(\lim_{V\to\infty}\frac{V}{4\pi^2}\right)\int_0^\infty dk\, k^2\omega_k\,. \tag{2.82}$$

Here, we see (i) a divergent *total energy* related to the fact that we are considering a homogeneous energy density in an infinite volume and (ii) a UV-divergent *energy density* for the vacuum; as ω_k behaves like $\sim k$ at the UV ($k^2 \gg m^2$), we see that the integrand grows cubically, which means the integral diverges quartically.

However, for a free theory in the absence of gravity, only energy differences are observable quantities, not absolute energy values. Thus, one can simply ignore this divergent energy value, by *redefining* the vacuum energy as 0. This can be sistematically achieved for any observables in the theory through the well-known procedure of *normal ordering*. In a normal-ordered observable, one just sets all annihilation operators to the right, and all creation operators to the left, so as not to have any residual contributions from the commutators; for example, we define the normal-ordered Hamiltonian:

$$:H: = \tfrac{1}{2}\sum_{\mathbf{k}}\omega_k(:a^\dagger_{\mathbf{k}}a_{\mathbf{k}}: + :a_{\mathbf{k}}a^\dagger_{\mathbf{k}}:) = \sum_{\mathbf{k}}\omega_k a^\dagger_{\mathbf{k}}a_{\mathbf{k}} = \sum_{\mathbf{k}}\omega_k N_{\mathbf{k}}. \tag{2.83}$$

Thus:

$$\langle 0|:H:|0\rangle = 0. \tag{2.84}$$

We note, however, that it is only in very special circumstances that this extremely simple procedure works for a physically meaningful cancelation of vacuum energy divergencies. In the next section, we shall see a less trivial example for which it no longer applies.

2.3 Vacuum Energy in Flat Space; the Casimir Effect

Much of the discussion in the present work regards vacuum energy. As we have seen in the last section, the most straightforward and naive approach to calculate it yields a divergent result. For free fields in Minkowski spacetime, this kind of divergency can be eliminated throughout by normal ordering, conventioning vacuum energy to be zero. In general, we will have to find a way to make sense of infinities which appear throughout for many observables through a systematic procedure of renormalization in curved spacetimes, which will be presented in more detail in Chapter 4. Still, even for free fields in flat spacetimes, one may find nontrivial vacuum effects, which cannot be accounted for by mere normal ordering. Thus, in this section, we shall make a preamble of the subject of renormalization, employing a simpler subtraction procedure to account for these nontrivial vacuum effects in flat spaces, and make a connection with one of the few instances where there are experimental results in the subject.

In the original Casimir effect, one explores the physical effect of vacuum energy for the electromagnetic field in the presence of two large parallel conducting plates, separated by small distance a, and grounded in a common potential. The situation is depicted as follows:

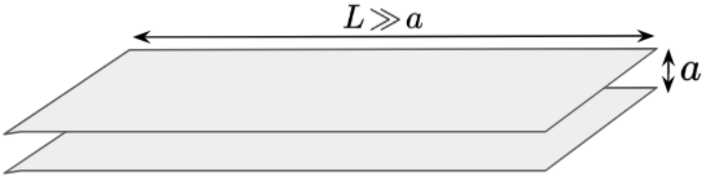

Figure 1 – Two parallel conducting plates, with characteristic width L, separated by a small distance $a \ll L$. Both plates are grounded at a common electric potential $V(z=0) = V(z=a) = cte$, usually conventioned to be 0.
Source: By the author.

The presence of these grounded conducting plates creates a nontrivial boundary condition for the electromagnetic field, which is forced to vanish on these surfaces. In practice, we want to analyze the behaviour of the field between the plates and far from their edges, so that we shall just take the continuum limit in both transverse directions $L \to \infty$. In such a situation, we could use the following field modes:

$$A^\mu \propto \epsilon^\mu e^{i\mathbf{k}_\perp \cdot \mathbf{x}_\perp} \sin(n\pi z/a) e^{-i\omega t} \tag{2.85}$$

(where A^μ is the usual electromagnetic 4-potential, and ϵ^μ represents a polarization vector).

In this setup, one can predict that there should be an attractive force between the plates due to vacuum fluctuation effects, shedding light on the nontrivial role that vacuum energy plays in Quantum Field Theory. This attractive force due to vacuum fluctuations between conductors is called the Casimir Effect; this effect has actually been measured in laboratory, making this one of the few instances of experimental evidence of vacuum energy (for a few historical details on the discovery and measurement of the Casimir Effect, check the last section of chapter 5 of (3), and references therein).

Inspired by this setup, we shall present here a simplified analogue model for the Casimir effect, using a massless scalar field with periodic boundary conditions. This model already allows us to compute a meaningful form of vacuum energy, and illustrates some of the general features of renormalized energy-momentum-stress observables, such as negative energy densities and pressures. Furthermore, we show that in this model it is more "energetically favorable" to have shorter period lengths a, mimicking the effects of an attractive force between plates in the actual Casimir setup.

Let us then develop the model in more detail. Since we chose periodic boundary conditions, we should work with modes of the form (2.34), whose allowed wave vectors are given by:

$$k^x, k^y = \frac{2\pi}{L} n^{x,y}, \qquad k^z = \frac{2\pi}{a} n^z, \qquad n^i \in \mathbb{Z}. \tag{2.86}$$

They correspond to the allowed frequencies:

$$\omega_k = |\mathbf{k}| = 2\pi \sqrt{\frac{i^2 + j^2}{L^2} + \frac{l^2}{a^2}}, \tag{2.87}$$

where we have denoted $n^{x,y,z}$ as i, j, l, respectively.

Now, the first step in our analysis of the vacuum energy of this field is to put a divergent expression like (2.82) in a *regularized* form. First, we write the energy density, which we shall regard as a function of the separation a, as *a limit of a convergent sum*:

$$\rho_0(a) = \frac{1}{V} \langle 0_a | H | 0_a \rangle = \frac{1}{2aL^2} \sum_{\mathbf{k}} \omega_k = -\frac{1}{2aL^2} \lim_{\alpha \to 0^+} \left[\frac{d}{d\alpha} \sum_{\mathbf{k}} e^{-\alpha \omega_k} \right]. \tag{2.88}$$

Note that for a finite α the exponential factor acts as a cutoff for arbitrarily high frequencies, taming the ultraviolet (UV), $|\mathbf{k}| \to \infty$, divergencies in the expression. Of course, we still get a divergent value for $\rho_0(a)$ when we take the limit $\alpha \to 0^+$. Our procedure then consists in keeping α temporarily finite – which is called *regularization* –, so that we can more closely identify the structures of the divergencies, and then find a meaningful physical subtraction to cancel them and obtain a finite result – which is called *renormalization*[P]. Once renormalization has been carried out, one may relax the regularization and take the limit $\alpha \to 0^+$, obtaining what is to be regarded as the physical prediction for that observable, to be compared with experiments[Q].

Now, the regularized expression for $\rho(a)$ reads:

$$\rho_0(\alpha, a) = -\frac{1}{2aL^2}\left[\frac{d}{d\alpha}\sum_{\mathbf{k}} e^{-\alpha\omega_k}\right], \tag{2.89}$$

for which we recover (2.88) as $\rho_0(a) = \lim_{\alpha\to 0^+}\rho_0(\alpha, a)$.

As our notation suggests, we are particularly interested in determining how the vacuum energy density depends on a, and whether we can separate a finite contribution from it out of this divergent expression. To achieve that, we are going to torture our regularized expression (2.89), making quite cumbersome operations on it, so that we may squeeze the a dependence out of the infinities.

First, we define an auxiliary function

$$S(\alpha, a) \equiv \frac{1}{L^2}\sum_{\mathbf{k}} e^{-\alpha\omega_k} \longrightarrow \frac{1}{(2\pi)^2}\sum_{l=-\infty}^{+\infty}\int d^2\mathbf{k}_\perp \exp\Big[-\alpha(\mathbf{k}_\perp^2 + (\tfrac{2\pi}{a})^2 l^2)^{1/2}\Big], \tag{2.90}$$

where we have taken the continuum limit ($L \to \infty$) for the transverse directions. Notice that, like ρ_0, this function's dependence on a is given implicitly by how it determines the domain Ω of allowed wave vectors $\mathbf{k}$ over which we perform the summation.

We can then write (2.90) as

$$\begin{aligned} S(\alpha, a) &= \frac{1}{(2\pi)^2}\sum_{l=-\infty}^{+\infty}\int d^2\mathbf{k}_\perp \exp\Big[-\alpha(\mathbf{k}_\perp^2 + (\tfrac{2\pi}{a})^2 l^2)^{1/2}\Big] \\ &= \frac{1}{2\pi}\int_0^\infty dk_\perp k_\perp e^{-\alpha k_\perp} + \frac{2}{2\pi}\sum_{l=1}^{\infty}\int_0^\infty dk_\perp k_\perp e^{-\alpha(k_\perp^2 + (\frac{2\pi}{a})^2 l^2)^{1/2}} \\ &= \frac{1}{2\pi}\Big[F(0) + 2\textstyle\sum_l F(l)\Big], \end{aligned} \tag{2.91}$$

where we have defined:

$$F(l) \equiv \int_0^\infty dk_\perp k_\perp e^{-\alpha\left[k_\perp^2 + \left(\frac{2\pi}{a}\right)^2 l^2\right]^{1/2}} = \left[\frac{1}{\alpha^2} + \frac{1}{\alpha}\frac{2\pi l}{a}\right] e^{-\frac{2\pi l}{a}\alpha}. \tag{2.92}$$

[P] The term *renormalization* is actually associated with a wider procedure in which one absorbs these subtracted infinities in a *redefinition* of basic parameters of the theory, such as masses and elementary charges. We shall discuss renormalization precedures in this more specific sense in chapter 4.3.

[Q] Sometimes comparisons of this type will involve the adjustment of one or several free renormalization parameters.

Since we are ultimately interested in taking the limit $\alpha \to 0^+$, we would like to write some kind of power expansion in α, so as to identify divergent, finite and vanishing terms. Here, it is very convenient to work with the Euler-Maclaurin formula for analytic functions:

$$\tfrac{1}{2}F(b) + \sum_{l=1}^{\infty} F(b+l) = \int_b^{\infty} dl F(l) - \sum_{m=1}^{\infty} \frac{B_{2m}}{(2m)!} F^{(2m-1)}(b), \tag{2.93}$$

being B_j the jth Bernoulli number (we have, for instance: $B_2 = 1/6$, $B_4 = -1/30$, $B_6 = 1/42$, etc.).

We now apply this formula to (2.91), setting $b = 0$. With a little algebraic effort, one may verify that:

$$F^{(1)}(0) = 0, \quad F^{(3)}(0) = 2\alpha\left(\frac{2\pi}{a}\right)^3, \quad \text{and} \quad F^{(j)}(0) = \mathcal{O}(\alpha^2), j \geq 5. \tag{2.94}$$

Besides, since $F(l)$ only depends on a through the combination l/a, we have that:

$$\int_0^{\infty} dl F(l) = aG(\alpha), \tag{2.95}$$

where $G(\alpha)$ does not depend on a. Thus, we may write:

$$\begin{aligned} \pi S(\alpha, a) &= aG(\alpha) - \frac{B_4}{24} F^{(3)}(0) + \mathcal{O}(\alpha^2) = aG(\alpha) + \frac{\pi^3}{45a^3}\alpha + \mathcal{O}(\alpha^2), \\ \Rightarrow \quad S(\alpha, a) &= \frac{a}{\pi} G(\alpha) + \frac{\pi^2}{45a^3}\alpha + \mathcal{O}(\alpha^2). \end{aligned} \tag{2.96}$$

Now, we substitute this result in our original expression for $\rho_0(a)$, (2.88). Note that when we take a derivative with respect to α and carry the limit $\alpha \to 0^+$ all of the terms $\mathcal{O}(\alpha^2)$ in (2.96) vanish, so that we are left with:

$$\rho_0(a) = -\frac{1}{2a} \lim_{\alpha \to 0^+} \left\{ \frac{d}{d\alpha} S(\alpha, a) \right\} = -\frac{1}{2\pi} \lim_{\alpha \to 0^+} G'(\alpha) - \frac{\pi^2}{90a^4} \tag{2.97}$$

(where the prime $'$ on G denotes its derivative with respect to α).

This expression, of course, still presents a divergence when we take the limit (we have just tortured our expression, we have not mutilated it yet). However, we have isolated this divergence in the first term, which does not depend on a. Now, since in flat spacetime we are not concerned with absolute values of the energy, but are rather interested in how it *varies* as we change the separation a, we have a freedom to redefine our 0-point energy and ignore this constant divergent term. An arguably natural choice for this 0-point is the Minkowski vacuum (corresponding to the limit $a \to \infty$); thus, we redefine the energy density through our regularized expression as:

$$\rho(a) \equiv \lim_{\alpha \to 0^+} \Big\{ \rho_0(\alpha, a) - \rho_0(\alpha, \infty) \Big\}. \tag{2.98}$$

Then, we are left just with the second term in (2.97):

$$\rho(a) = -\frac{\pi^2}{90a^4}, \tag{2.99}$$

which makes for a quite nice expression after those lengthy calculations. Torture is over.

Let us now take a moment to analyze and interpret our results physically. First, note that the energy density we obtained is *negative*. Of course, this is only necessarily so because we have defined the reference Minkowski vaccum energy as 0; we could well add any constant C to that density ($\rho \to \rho + C$), and obtain a (possibly positive[R]) physically equivalent result in flat spaces. Second, note that ρ *decreases as* a *decreases*, regardless of the choice of C (for $C = 0$, it becomes *more negative*). To look at the consequences of that more closely, it is clarifying to look at the transverse energy surface density, $\sigma = a \times \rho$, or, more conveniently, return to a finite *fixed* transverse size, $L \gg a$, which yields a finite *total* vacuum energy:

$$E(a) = aL^2 \times \rho(a) = -\frac{\pi^2}{90a^3}L^2 \qquad \Rightarrow \qquad \sigma(a) = -\frac{\pi^2}{90a^3}. \tag{2.100}$$

We see that E also *decreases* with a, so that it should be more "energetically favorable" to have arbitrarily small a values. Analyzing the internal work that would be necessary to vary/expand a, we obtain a negative pressure p, given by:

$$dE = -pdV \qquad \Rightarrow \qquad p(a) = \frac{1}{L^2}\frac{dE}{da} = -\frac{\pi^2}{30a^4}. \tag{2.101}$$

We emphasize that the pressure (2.101) turns out negative regardless of the choice of C for the energy density. Of course, in our periodic condition setup there is no physical boundary to move, and to empirically analyze this pressure; it merely reflects the theoretical exercise of varying an arbitrary field periodicity length. However, we shall find an entirely analogous result in the original Casimir setup, where there is a physical boundary given by the conducting plates.

[R] However, since (2.99) is unbounded as we make a arbitrarily small, we can only have a positive ρ for *any* values of a by making C infinite.

In the electromagnetic case, there will be just two relevant differences in the calculation: the modes will be of the form (2.85), rather than (2.34), and there are two independent polarizations for the electromagnetic field. The latter will simply give us a 2 factor, whereas the former has the effects of (i) changing the allowed k^z vectors $2\pi l/a \to \pi l/a$ – this will inflict a change in the value of $F^{(3)}$, $2\alpha(\frac{2\pi}{a})^3 \to 2\alpha(\frac{\pi}{a})^3$ –, and (ii) making the modes $+l$ and $-l$ linearly dependent (note that the $l = 0$ mode makes no contribution to the renormalized ρ). Summarizing these factors, we obtain:

$$\rho_{EM}(a) = -\frac{1}{8} \times \frac{1}{2} \times 2 \times \frac{\pi^2}{90a^4} = -\frac{\pi^2}{720a^4}, \tag{2.102}$$

$$p_{EM}(a) = \frac{1}{L^2}\frac{dE_{EM}}{da} = -\frac{\pi^2}{240a^4}. \tag{2.103}$$

In flat space, one often interprets these negative values as being merely relative to the 'outside region' – an embedding Minkowski spacetime –, since there are less modes that "fit" in the finite length a (imposing the appropriate boundary conditions). Furthermore, this explanation seems quite natural in both our simplified periodic setup and the original Casimir one, as both can be embedded in a larger, Minkowski space. In the latter, then, one interprets the attractive force between the plates as being due to a higher *positive vacuum pressure outside, pushing the plates together.*

However, the situation is radically different in curved spacetimes, where this kind of interpretation is no longer generally attainable, for two basic reasons. First, when we take gravity into account, we must ascribe physical meaning to *absolute values* of energy (more precisely of energy-momentum-stress), which act as the source of curvature in Einstein's Equations, so that one is no longer at liberty of considering only energy differences. Secondly, there is generally no embedding spacetime (or 'outside' region) to compare to, so that one is forced to analyze the quantum fluctuation effects of renormalized observables *intrinsically.* In particular, this means that we can end up with physical, renormalized negative energy densities and pressures, even when those are classically positive-definite.

2.4 Formal Remarks on Expansions in Normal Modes

Already from the fact that there are divergencies in the theory, we can anticipate that there are formal issues not fully addressed in the presentation so far (indeed, one can glimpse in Appendix A that these emerge from a forced attempt to make sense of products of distributions). Although the present text does not aim at providing a fully rigorous treatment of QFT, it is the author's personal belief that a more thorough presentation of some of its formal aspects may be very enlightening both operationally and conceptually, especially when we must handle intricate and often physically nebulous topics such as renormalization. A more complete treatment of the topics covered in this and the next section is given in (3) (chapters 2-4), from where most of the exposition here is based and supplementary material to this discussion can be found in Appendix A.

The problem of expansion in normal modes is one of Linear Algebra. Given a Hilbert space that contains all the acceptable physical states in our theory (and excludes all the unphysical ones) – armed with a complete set of elementary observables which allows us to build any physical observable to probe it –, we typically want a convenient basis in terms of which we may expand any state in it. While this is a relatively trivial task for any finite-dimensional Hilbert space, it involves some subtleties when we go to infinite dimensions.

Already in more elementary instances such as nonrelativistic Quantum Mechanics, one is faced with the problem of nonnormalizable wave functions in the continuum. In this context, one usually breaks the types of eigenvalue problem in two instances:

1 - Discrete spectrum $\{E_j\}$: in this case, one can find a complete set of normalizable wave functions ψ_j, which form an ordinary orthornormal basis in the Hilbert space $\mathcal{H} \subset \mathcal{L}^2$ of the theory:

$$H\psi_j = E_j\psi_j, \quad \langle \psi_j, \psi_k \rangle = \delta_{jk}. \tag{2.104}$$

In this instance, one can expand any wave function ϕ in terms of this basis:

$$\phi(j) \equiv \langle \psi_j, \phi \rangle \quad \Rightarrow \quad \phi(x) = \sum_j \phi(j)\psi_j(x), \tag{2.105}$$

as well as easily evaluate the action of operators:

$$H\phi = \sum_j E_j \phi(j)\psi_j. \tag{2.106}$$

2 - Continuous spectrum $\{E_\lambda\}$: in this case, the functional solutions f_λ to the differential equation:

$$Hf = E_\lambda f_\lambda \quad \Leftrightarrow \quad (H - E_\lambda)f_\lambda = 0 \tag{2.107}$$

will not generally belong to the Hilbert space $\mathcal{H}$, and therefore will not be proper wave functions. Nevertheless, the Spectral Theorem (see chapter 2 of (3)) assures that one may still project any wave function $\phi \in \mathcal{H}$ in all of these modes and write the expansions:

$$\phi(\lambda) = \langle f_\lambda, \phi \rangle \quad \Rightarrow \quad \phi(x) = \int_{\sigma(H)} d\lambda\, \phi(\lambda) f_\lambda(x), \tag{2.108}$$

as well as:

$$H\phi = \int_{\sigma(H)} d\lambda\, \phi(\lambda) E_\lambda f_\lambda. \tag{2.109}$$

So far, we have treated both the discrete index j and the continuous one λ as nondegenerate. It may well happen that there are degeneracies. For instance, in the continuous case, we could have a free particle with a nonzero spin, which would oblige us to modify an expression like (2.108) to:

$$\phi_i(\lambda) \equiv \langle f_{i,\lambda}, \phi\rangle \quad \Rightarrow \quad \phi(x) = \int_{\sigma(H)} \sum_i \phi_i(\lambda) f_{i,\lambda}(x) d\lambda. \tag{2.110}$$

More generally, there may also be continuous degeneracies in each eigenvalue, and/or the dimension of each subspace may depend on the eingenvalue E. One may even have a mixture of the two cases considered above (as it happens in hydrogen atom, where one has a point spectrum for bounded states $E < 0$ and a continuous spectrum for unbounded states $E > 0$). To avoid more cumbersome notations and the need to split between various cases, we condense our notation through a single, nondegenarate index λ (λ may belong to a multidimensional space, and comprise both continuous and discrete indices) and write our expansions as:

$$\phi(x) = \int_{\sigma(H)} d\mu(\lambda)\, \langle u_\lambda, \phi\rangle\, u_\lambda(x)\,, \tag{2.111}$$

where $\mu(\lambda)$ represents a measure over the spectrum[S]. In continuous portions of $\sigma(H)$, $\mu(\lambda)$ will be a continuous, monotonically increasing function, whereas in the discrete portions, it will be a constant function with discontinuous "jumps" at $\lambda \in \sigma(H)$ (so that $d\mu(\lambda)$ will be a countable sum of Dirac deltas).

Having introduced this unified notation, we take the chance to explore slightly more general linear scalar field equations, in the form:

$$-\frac{\partial}{\partial t^2}\phi(\mathbf{x}, t) = H^2_{\mathbf{x}}\phi(\mathbf{x}, t) \qquad \Leftrightarrow \qquad \left[\partial_t^2 + H^2_{\mathbf{x}}\right]\phi(\mathbf{x}, t) = 0, \tag{2.112}$$

where $H_{\mathbf{x}}$ is an elliptic differential operator[T] acting only in the spatial variables (we recover the familiar Klein-Gordon equation (2.30) by making $H^2_{\mathbf{x}} = -\nabla^2_{\mathbf{x}} + m^2$). Equation (2.112) still allows for a mode decomposition in the form:

[S] If the reader is unfamiliar with the concept of a measure, we recommend section 1.D of (19).

[T] For practical purposes, one must not worry with the precise definition of an elliptic operator; here, it will be a technical requirement for the Hamiltonian to be bounded from below, so that there is a stable vacuum state. The reader will find more precise definitions and thorough discussion in (3).

$$u_\lambda(\mathbf{x},t) = \frac{e^{-i\omega_\lambda t}}{\sqrt{2\omega_\lambda}}\psi_\lambda(\mathbf{x}), \tag{2.113}$$

where:

$$H_{\mathbf{x}}^2\psi_\lambda(\mathbf{x}) = \omega_\lambda^2\psi_\lambda(\mathbf{x}). \tag{2.114}$$

And therefore, we write an expansion for ϕ in the form:

$$\phi(x) = \int_{\sigma(H)} d\mu(\lambda)\, a(\lambda)u_\lambda(x) + a^\dagger(\lambda)u_\lambda^*(x), \tag{2.115}$$

for which the commutation relations read:

$$[a(\lambda), a(\lambda')] = [a^\dagger(\lambda), a^\dagger(\lambda')] = 0, \tag{2.116a}$$

$$[a(\lambda), a^\dagger(\lambda')] = \delta(\lambda, \lambda'), \tag{2.116b}$$

where $\delta(\lambda, \lambda')$ is the delta distribution with respect to the measure μ, that is:

$$\int_{\sigma(H)} d\mu(\lambda')f(\lambda')\delta(\lambda,\lambda') = f(\lambda). \tag{2.117}$$

We shall make use of this more general, unified expansion in the next section, where we analyze integral kernels to the wave equation (2.112) and two-point functions.

We shall make use of this more general, unified expansion in the next section, where we analyze integral kernels to the wave equation (2.112) and two-point functions.

Also, we note that in the continuum case there will be a subtlety in the definition of field modes occupations and Fock space. Just as eigenvectors from positions and momenta, the states determined by application of one creation operator in the continuum, $|1_\lambda\rangle = a^\dagger(\lambda)\,|0\rangle$, must be understood in a generalized, distributional sense (see Appendix A). Actual one-particle states will be given by integrals of these generalized states in a continuous interval:

$$|\psi_1\rangle = \int_{\sigma(H)} d\mu(\lambda')\rho(\lambda)a^\dagger(\lambda)\,|0\rangle\,, \tag{2.118}$$

and one can similarly write n-particle states $|\psi_n\rangle$ with integrals of products of n creation operators.

Throughout most of this dissertation, however, we shall simply write our states as discrete sums with the usual notation in Fock spaces, and the appropriate generalization will be implied in them continuum.

2.5 Two-point Functions

To conclude this chapter, we give an overview of a very important class of functions, useful to perform many computations in the theory: two-point functions. These are the (number-valued) expectation values of observables bilinear in field amplitudes at two spacetime events[U] x and x':

$$G(x,x') = \langle\Psi|f\Big(\phi(x),\phi(x')\Big)|\Psi\rangle\,, \tag{2.119}$$

where $f\Big(\lambda\phi(x),\lambda'\phi(x')\Big) = \lambda\lambda' f\Big(\phi(x),\phi(x')\Big)$.

These will be central in computing very important physical quantities, such as field correlations $f=\phi(x)\phi(x')$, commutators $f=[\phi(x),\phi(x')]$ and anticommutators $f=\{\phi(x),\phi(x')\}$. Since many of these bilinear observables are actually proportional to the identity operator (as is the case with the commutator of scalar fields), their corresponding two-point functions will actually be state-indepent. Generally, however, they may bear a state dependence (which is quite natural, for instance, for field correlations), and a state that will have central importance for computing them is the vacuum $|0\rangle$. In fact, it turns out that vacuum expectation values of various of these bilinear observables can be identified with various Green Functions of the field equations.

Turning to the example of the more general scalar field introduced in the previous section, we then begin our analysis by investigating the Green functions of the wave equation (2.112):

$$\Big[\partial_t^2 + H_{\mathbf{x}}^2\Big]G(x,x') = \delta(x,x') = \delta(t-t')\delta(\mathbf{x}-\mathbf{x}'), \tag{2.120}$$

where we have made explicit use of global inertial coordinates to split $\delta(x,x')$ into spatial and temporal Dirac deltas[V].

This integral kernel (Green function) allows us to write the classical field solutions $g(x)$ to our wave equations with a source $J(x)$, $[\partial_t^2 + H_{\mathbf{x}}^2]g(x) = J(x)$, in the form:

$$g(x) = \int d^4x'\,J(x')G(x,x') + \text{Homogeneous solution}. \tag{2.121}$$

If we then take a formal Fourier transform in time t and a spectral transform in space $\mathbf{x}$ ($\int dt e^{i\omega t}\int d^3\mathbf{x}\,\psi_{\lambda}^{*}(\mathbf{x})$) in equation (2.120), we obtain:

[U] As the reader may have noted from appendix A, these are generally not functions, but distributions. Still, we maintain the terminology throughout the section and the rest of the dissertation.

[V] We just need a weaker split between space and time, but we will avoid getting more technical at this point.

$$(-\omega^2 + \omega_\lambda^2)\tilde{G}(\omega, \lambda; x') = e^{i\omega t'}\psi_\lambda^*(x'). \tag{2.122}$$

Then, reversing this formal Fourier transform, we obtain the following expansion for $G(x, x')$:

$$G(x, x') = -\frac{1}{2\pi}\int d\mu(\lambda)\int d\omega \frac{1}{\omega^2 - \omega_\lambda^2} e^{-i\omega(t-t')}\psi_\lambda^*(\mathbf{x}')\psi_\lambda(\mathbf{x})\,. \tag{2.123}$$

When we attempt to make sense of this formal expression, starting from the ω integral in the right, we run into trouble due to the poles of the integrand at $\pm\omega_\lambda$. How, then, should we compute (and interpret) this expression? Well, as we are carrying this integral in the real axis, one could propose we displace the poles by a distance ϵ in the complex plane, carry on the finite (regularized) integral, and then try to take the limit $\epsilon \to 0$. Equivalently, one could leave the poles fixed and displace the integration contour a little around them[W]. As there are numerous ways to displace the contours, we end up with correspondingly numerous Green functions.

Before we actually carry the various integrations, it is worth noting explicitly that, for $t > t'$, $e^{-i\omega(t-t')}$ will exponentially diverge in the upper half complex plane ("as $\omega \to +i\infty$") and exponentially decay in the lower half complex plane ("as $\omega \to -i\infty$"), and vice-versa for $t < t'$. To assure the contributions outside the real axis will vanish, we *should always close the contour in the decaying region* (*e.g.* by a semicircle at infinity), so that the encompassed poles will depend on the considered times. With those considerations, let us enumerate and compute a few relevant Green function, giving the prescription for their respective contours:

1- Retarded Green Function, G_{ret}**:** pass the contour *above* both poles (see Figure 2), yielding:

$$G_{ret}(x, x') = \begin{cases} 0, & t < t' \qquad \text{(2.124a)} \\ -2\pi i\Big(\mathrm{Res}_I(\omega\!=\!+\omega_\lambda) + \mathrm{Res}_I(\omega\!=\!-\omega_\lambda)\Big), & t > t' \qquad \text{(2.124b)} \end{cases} .$$

[W] In this case, one does not have to actually take any limits to bring the contour back into the real axis, as the integral will be evaluated by Cauchy's theorem, remaining invariant unless the contour crosses a pole. If the reader is not particularly comfortable with contour integrations, we recommend a brief review, *e.g.*, in chapters 6 and 7 of (20).

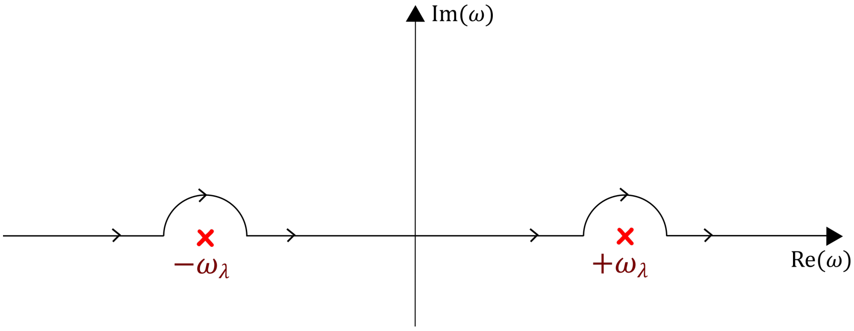

Figure 2 – Contour for the Retarded Green Function. As the other contours, it must be closed at the upper half of the complex plane for $t < t'$ and at the lower half for $t > t'$.
Source: By the author.

Thus, for $t > t'$:

$$\begin{aligned} G_{ret}(x,x') &= \int d\mu(\lambda) \frac{i}{2\omega_\lambda} \Big[e^{-i\omega_\lambda (t-t')} - e^{+i\omega_\lambda (t-t')} \Big] \psi^*_\lambda(\mathbf{x}')\, \psi_\lambda(\mathbf{x}) \\ &= \int \frac{d\mu(\lambda)}{\omega_\lambda} \sin\Big(\omega_\lambda (t-t')\Big) \psi^*_\lambda(\mathbf{x}')\, \psi_\lambda(\mathbf{x}). \end{aligned} \tag{2.125}$$

We can then see that, in eq (2.121), this kernel would correspond to a solution $g(x)$ of the field equations that incorporates the source $J(x')$ only to the past of x (we shall see ahead that it is actually supported in the past light cone of x, which will independ of the arbitrary choice of simultaneity for space separated events).

2- Advanced Green Function, G_{adv}: pass the contour *under* both poles (see Figure 3), so that:

$$G_{ret}(x,x') = \begin{cases} +2\pi i \Big(\mathrm{Res}_I(\omega = +\omega_\lambda) + \mathrm{Res}_I(\omega = -\omega_\lambda) \Big), & t < t' \quad (2.126a) \\ 0, & t > t' \quad (2.126b) \end{cases} .$$

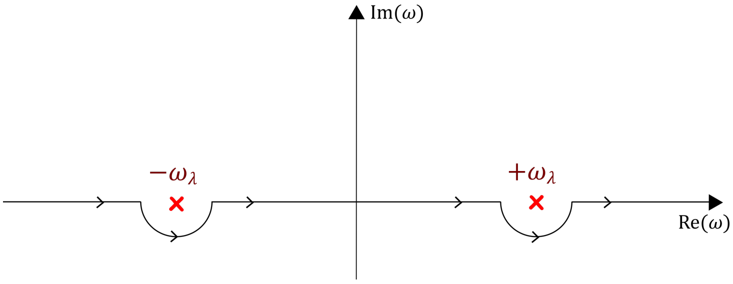

Figure 3 – Contour for the Advanced Green Function.
Source: By the author.

Then, similarly to (2.125), we have for $t < t'$:

$$G_{adv}(x,x') = -\int \frac{d\mu(\lambda)}{\omega_\lambda} \sin\Big(\omega_\lambda(t-t')\Big)\psi^*_\lambda(\mathbf{x}')\,\psi_\lambda(\mathbf{x}). \tag{2.127}$$

Correspondingly, this kernel yields a solution that incorporates the source $J(x')$ only to the future of x (complementary to G_{ret}, G_{adv} will only be supported in the future light cone of x).

3- The Feynman propagator, G_F**:** go *under the left pole* ($\omega = -\omega_\lambda$), and *over the right one* ($\omega = +\omega_\lambda$), so that one or the other will contribute for $t > t'$ or $t < t'$.

$$G_F(x,x') = \begin{cases} +2\pi i\,\mathrm{Res}_I(\omega{=}-\omega_\lambda), & t<t' \qquad (2.128a) \\ -2\pi i\,\mathrm{Res}_I(\omega{=}+\omega_\lambda), & t>t' \qquad (2.128b) \end{cases}.$$

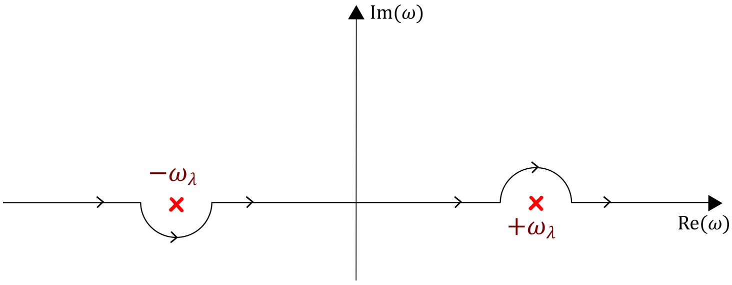

Figure 4 – Contour for the Feynman propagator.
Source: By the author.

Carrying the residue integration, we find:

$$G_F(x,x') = i\int \frac{d\mu(\lambda)}{2\omega_\lambda} e^{-i\omega_\lambda|t-t'|}\psi_\lambda(\mathbf{x}')\,\psi^*_\lambda(\mathbf{x}). \tag{2.129}$$

This will be a crucial Green Function, as it represents a particularly important integral kernel to the inverse of our differential operator, $[\partial_t^2 + H_\mathbf{x}]^{-1}$, as stated in equation (2.120). To see this more clearly, we cast the residue integral in a different form, displacing both poles an infinitesimal distance from the real axis, as (see Figure 5):

$$\begin{cases} \omega = -\omega_\lambda \\ \\ \omega = +\omega_\lambda \end{cases} \qquad \longrightarrow \qquad \begin{cases} \omega = -\omega_\lambda + i\epsilon \\ \\ \omega = +\omega_\lambda - i\epsilon \end{cases},$$

so that, to first order in ϵ:

$$\omega^2 + \omega_\lambda(-\omega_\lambda) \qquad \longrightarrow \qquad \omega^2 + (\omega_\lambda - i\epsilon)(-\omega_\lambda + i\epsilon) = \omega^2 - \omega_\lambda^2 + i\epsilon \tag{2.130}$$

(where we have absorbed a positive factor $\omega_\lambda/2$ into ϵ in the last equality).

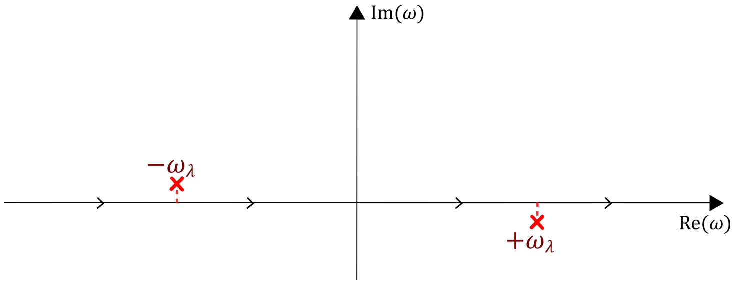

Figure 5 – Contour for the Feynman propagator in the real axis, with the poles correspondingly displaced in the imaginary plane.
Source: By the author.

Thus, we write G_F as:

$$G_F(x,x') = \lim_{\epsilon\to 0^+} \left\{ -\frac{1}{2\pi} \int d\mu(\lambda) \int d\omega \frac{e^{-i\omega_\lambda(t-t')}}{\omega^2 - \omega_\lambda^2 + i\epsilon} \psi^*_\lambda(\mathbf{x}')\, \psi_\lambda(\mathbf{x}) \right\}. \tag{2.131}$$

(Note that, using (A.18) – and paying proper attention to which pole we are encompassing for each t – one can easily recover (2.129).)

From this form, it is easy to see that it will correspond to the following integral kernel (again, working in first order in ϵ):

$$G_F = \lim_{\epsilon\to 0^+} \left\{ \mathrm{Ker}\Big([\partial_t^2 + H_{\mathbf{x}}^2 - i\epsilon]^{-1} \Big) \right\}. \tag{2.132}$$

Indeed, in chapter 4 we will make use of it in the renormalization of the effective action. Also, this $i\epsilon$ factor will play the role of a regularizer for path integrals, making them well defined when we vary e^{iS} taking field amplitudes up to infinity.

Finally, we note that this integral kernel will obey a Green equation *with a reversed sign*:

$$[\partial_t^2 + H_{\mathbf{x}}^2] G_F(x,x') = -\delta(x,x') = -\delta(t-t')\delta(\mathbf{x}-\mathbf{x}') \tag{2.133}$$

4- Principal Value Green Function, $\bar{G}$: pass the integration contour directly through the poles, taking the *principal value* at each one (see Appendix A for more details on principal-value distributions).

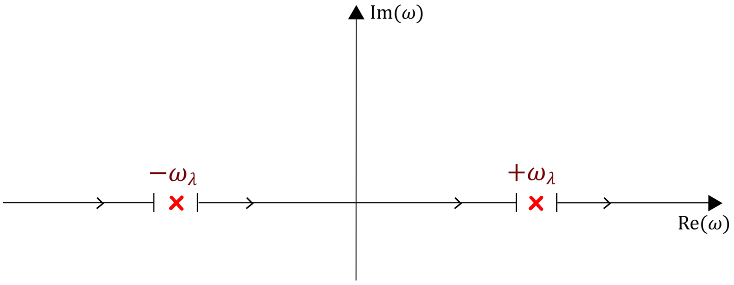

Figure 6 – Integration representation for the Principal Value Green Function. Here, one must approach each pole symmetrically from both sides, to cancel out the divergent terms.
Source: By the author.

This contour can be thought of as the juxtaposition of the advanced and retarded contours (more precisely, half of this jusxtaposition, so one does not count the integral twice), as we try to sketch in Figure 7. Thus, we have:

$$\bar{G}(x,x') \equiv \frac{1}{2}\Big(G_{ret}(x,x') + G_{adv}(x,x')\Big). \tag{2.134}$$

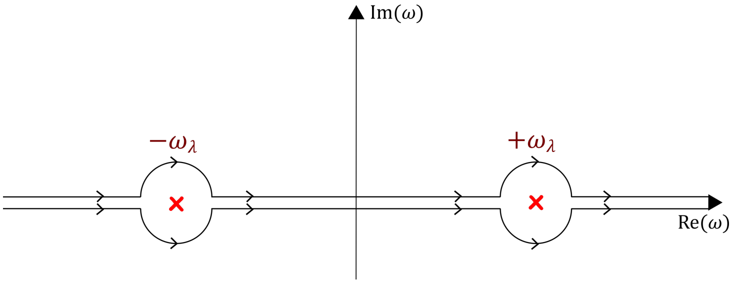

Figure 7 – The Principal Value contour represented as (half of) the juxtaposition of the advanced and retarded contours.
Source: By the author.

All of the kernels presented so far are *actual Green functions*, obeying the inhomogeneous equation (2.120). By taking the difference between pairs of them, we arrive at solutions to the *homogeneous* wave equation, corresponding to closed contours in the complex plane. In the literature, these are also referred to as Green functions, so we shall maintain that terminology. We enumerate a few relevant ones:

5- Wightman Function $-iG^+$: take a retarded contour and subtract a (properly adjusted) Feynman one, so that we end up with a closed curve around the right pole (see Figure 8). This yields:

$$G^+ = \int \frac{d\mu(\lambda)}{2\omega_\lambda} e^{-i\omega_\lambda(t-t')} \psi^*_\lambda(\mathbf{x}')\, \psi_\lambda(\mathbf{x}). \tag{2.135}$$

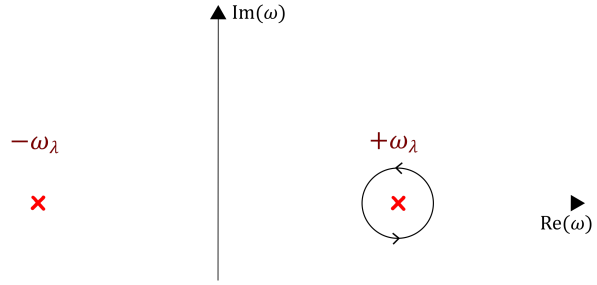

Figure 8 – Closed contour corresponding $-iG^+$. Here, one goes around the right pole once, counterclockwise.
Source: By the author.

6- Wightman Function $+iG^-$: similarly by taking the Feynman contour and subtracting the advanced one, we encompass the left pole, yielding:

$$G^- = \int \frac{d\mu(\lambda)}{2\omega_\lambda} e^{+i\omega_\lambda(t-t')} \psi^*_\lambda(\mathbf{x}')\, \psi_\lambda(\mathbf{x}). \tag{2.136}$$

We also note here that $G^- = (G^+)^*$.

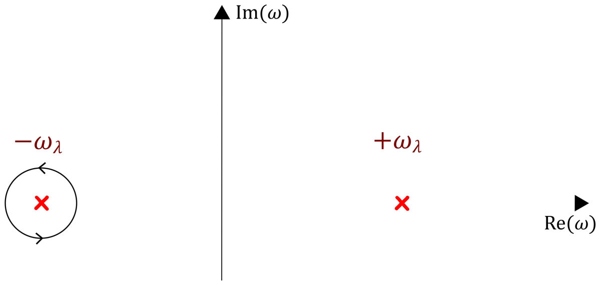

Figure 9 – Closed contour corresponding iG^-. Here, one goes around the left pole once, counterclockwise.
Source: By the author.

7-The Commutator G: go around both poles counterclockwise. This can be obtained subtracting G_{adv} from G_{ret}:

$$G = G_{adv} - G_{ret} = -i(G_+ - G_-) = 2\,\mathrm{Im}(G_+). \tag{2.137}$$

(The reason why we call this function the commutator will be clear briefly, when we analyze the connection with field operators.) We have its functional form from (2.137):

$$G(x,x') = -\int \frac{d\mu(\lambda)}{\omega_\lambda} \sin\Big(\omega_\lambda(t-t')\Big)\psi^*_\lambda(\mathbf{x}')\, \psi_\lambda(\mathbf{x}). \tag{2.138}$$

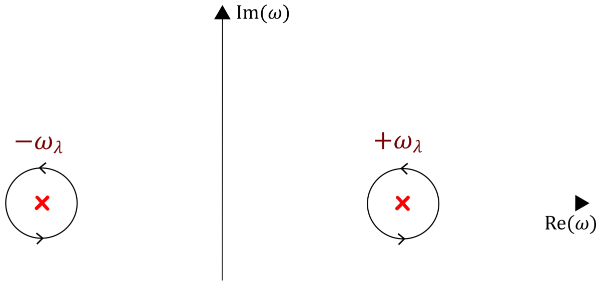

Figure 10 – Closed contour corresponding G. Here, one goes around both poles, counterclockwise.
Source: By the author.

8- The Anticommutator (also known as Hadamart elementary function or Schwinger function) $G^{(1)}$: go around the right pole clockwise, and the left one counterclockwise.

$$G^{(1)} = G^{+} + G^{-} = 2\,\mathrm{Re}(G_{+})$$
$$= -i[(G_F - G_{adv}) + (G_F - G_{ret})] = 2i(\bar{G} - G_F). \tag{2.139}$$

Thus:

$$G^{(1)}(x, x') = \int \frac{d\mu(\lambda)}{\omega_\lambda} \cos\Big(\omega_\lambda (t - t')\Big) \psi^{*}_{\lambda}(\mathbf{x}')\, \psi_\lambda(\mathbf{x}). \tag{2.140}$$

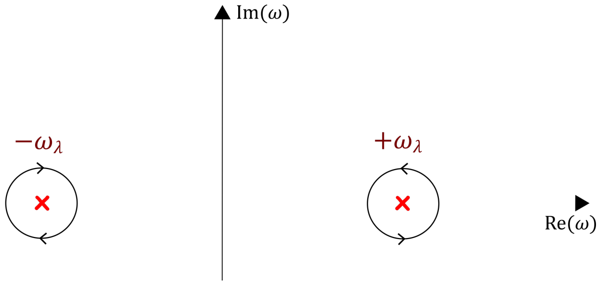

Figure 11 – Closed contour corresponding $G^{(1)}$. One goes around the right pole clockwise and the left one counterclockwise.
Source: By the author.

One important feature to notice in these Green functions concerns their spacetime support. Taking a fixed x, we have already seen that G_{ret} and G_{adv} are null for $t < t'$ and $t > t'$, respectively. To take the analysis further, it is particularly enlightening to consider $t = t'$. We see that, at equal times, G, G_{ret}, G_{adv}, $\bar{G}$ all vanish (whereas $G^{\pm}$, G_F and $G^{(1)}$ do not, even for $\mathbf{x} \neq \mathbf{x}'$). Further, by analyzing first time derivative of G, for example, we find:

$$\partial_t G(x,x')\Big|_{t=t'} = -\int d\mu_\lambda \cos\Big(\omega_\lambda \times 0\Big)\psi^*_\lambda(\mathbf{x}')\,\psi_\lambda(\mathbf{x}) = \delta(\mathbf{x}-\mathbf{x}'). \tag{2.141}$$

Due to this localized initial data, G (as a function of x') will be only supported inside the light cone of x. Consequently (see eq. (2.137)), G_{adv} will be only supported in the future light cone of x, and G_{ret} in its past light cone. In contrast, $G^{\pm}$ and $G^{(1)}$ spread through all spacetime, even for spacelike separated events.

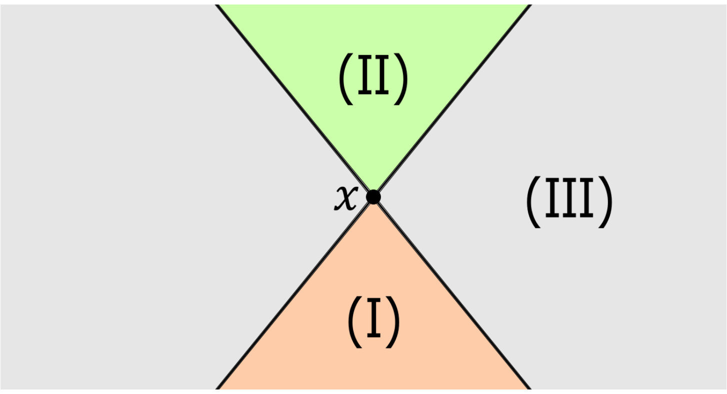

Figure 12 – Green functions and their various supporting regions. G_{ret} will only be nonzero inside the past light cone (region (I)), G_{adv}, inside the future light cone, and G, $\bar{G}$ inside both the entire light cone (regions (I) and (II)). On the other hand, $G^{\pm}$ and $G^{(1)}$ will also be nonvanishing in the spacelike-separated region (III), spreading through all of spacetime.
Source: By the author.

Now that we have constructed various Green functions as solutions of the field equation, we shall identify them with the correspondent vacuum expectation values of bilinear field operators. When we make the formal expansion of a few of the vacuum expectation values mentioned above, we see that we can immediately identify them with some of the above Green functions. For example, let us evaluate the product $\phi(x)\phi(x')$:

$$\begin{aligned}\langle 0|\phi(x)\phi(x')|0\rangle &= \iint \frac{d\mu(\lambda)}{\sqrt{2\omega_\lambda}}\frac{d\mu(\lambda')}{\sqrt{2\omega_{\lambda'}}}\,\langle 0\Big|\Big(a_\lambda u_\lambda(x) + a^\dagger_\lambda u^*_\lambda(x)\Big)\Big(a_{\lambda'}u_\lambda(x') + a^\dagger_{\lambda'}u^*_{\lambda'}(x')\Big)\Big|0\rangle \\ &= \iint \frac{d\mu(\lambda)}{\sqrt{2\omega_\lambda}}\frac{d\mu(\lambda')}{\sqrt{2\omega_{\lambda'}}} u_\lambda(x)u^*_{\lambda'}(x')\delta(\lambda,\lambda') \\ &= \int \frac{d\mu(\lambda)}{2\omega_\lambda} e^{-i\omega_\lambda(t-t')}\psi_\lambda(\mathbf{x})\psi^*_\lambda(\mathbf{x}').\end{aligned} \tag{2.142}$$

Thus, we immediately identify it with the Wightman function:

$$G^+(x,x') = \langle 0|\phi(x)\phi(x')|0\rangle\,. \tag{2.143}$$

Similarly, $\phi(x')\phi(x)$ yields:

$$G^{-}(x,x') = \langle 0|\phi(x')\phi(x)|0\rangle\,. \tag{2.144}$$

Then, from eqs (2.137) and (2.139), we immediately identify the commutator and anticommutator:

$$iG(x,x') = \langle 0|[\phi(x)\phi(x')]|0\rangle\,, \tag{2.145}$$

$$G^{(1)}(x,x') = \langle 0|\{\phi(x)\phi(x')\}|0\rangle\,. \tag{2.146}$$

Note that this last function symmetrizes the two-point field product before evaluating its expectation value. For this reason, one often computes $G^{(1)}$, instead of working with G^{+} and/or G^{-} directly.

Finally, we identify Feynman's propagator with the time-ordered product:

$$-iG_F(x,x') = \langle 0|\mathcal{T}(\phi(x)\phi(x'))|0\rangle\,, \tag{2.147}$$

where we have defined:

$$\mathcal{T}(\phi(x)\phi(x')) = \phi(x)\phi(x')\Theta(t-t') + \phi(x')\phi(x)\Theta(t'-t) = \begin{cases}\phi(x)\phi(x'), & t > t' \\ \phi(x')\phi(x), & t' > t\end{cases}\,. \tag{2.148}$$

Having reconstructed all of these operators as expectation values, we shall not rederive that they obey the (homogeneous or inhomogeneous) wave equations in all cases. We just note that, taking in consideration that $\phi(x)$ obeys the homogeneous field equation (2.112) (and $[\partial_t^2 + H_{\mathbf{x}}^2]$ does not act on x'), and that:

$$\partial_t\Theta(t-t') = \delta(t-t'), \qquad \int d\mu(\lambda)\psi(\mathbf{x})\psi^*(\mathbf{x}') = \delta(\mathbf{x}-\mathbf{x}'), \tag{2.149}$$

it is straightforward to recover them from the field-operator definition.

3 QUANTUM FIELD THEORY IN CURVED SPACETIME

In this chapter, we shall generalize in a straightforward manner the basic formalism of quantization of noninteracting fields to curved spacetimes, explicitly developed through the paradigmatic example of a real scalar field. In this particular quantization procedure, we promptly use the existence of a decomposition of the solutions of the classical field equations in orthonormal modes to impose the usual commutation relations $[a_i, a_j^\dagger] = \delta_{ij}$ and follow in general lines some of its consequences.

To concretely carry this procedure, we start by defining the classical prerequisites for our theory in curved spacetime. In section 3.1, we give the basic outline of theory of General Relativity and how to formulate the joint dynamics of matter and spacetime in a Lagrangian formalism. We also generalize the notion of 'equal-time' surfaces to (*globally hyperbolic*) curved spacetimes, defining the notion of *Cauchy Surfaces*.

Then, in section 3.2, we explicitly develop the quantization procedure for a noninteracting scalar field in curved spaces, and discuss some basic aspects of QFTCS, such as the absence of a physically priviledged vacuum state, as well as how it requires bosonic statistics (commutation relations) to be internally consistent under general mode transformations.

After that, in section 3.3, we give particular emphasis to the construction of an operational definition of particles, based on the response of actual particle detectors, as well as to the nontrivial relation between the different vacua associated with different mode decompositions, which prepares the ground for a more meaningful discussion of the processes of particle creation in dynamical spacetimes in section 3.4.

Finally, in section 3.5, we analyze the limits in which descriptions in terms of particle modes are meaningful to define what is called the adiabatic vacuum in dynamical spacetimes. Along with the corresponding asymptotic expansions of the field modes (the so-called adiabatic expansions), it will play a key role in the discussion of renormalization in the following chapter.

3.1 General Relativity and the Structure of Spacetime

In the present section, we give a brief overview of General Relativity (GR), discussing some of its geometrical and dynamical features. After laying key aspects in the interplay between matter (fields) and spacetime geometry given by the Einstein Equations, we show how these can be derived through a minimal action principle, extending the formalism of section 2.1 that will allow for Lagrangian formulation of GR permeated by matter fields. Given the overwhelming challenges in obtaining a fully quantum theory of gravity (15) (either in the vacuum or in the presence of matter), we turn to the well-established and fruitful approach of quantizing matter fields in a classical curved spacetime.

theory of General Relativity is without doubt a major revolution in the way we conceive space, time, and gravity. Rather than a static immutable stage through which matter propagates passively, spacetime comes to be conceived as a dynamic entity, curved by the matter within it. In a formal perspective, this step is achieved by letting go the assumption that spacetime is decribed as a flat space armed with a *given* flat (Minkowski) metric $(\mathbb{R}^4, \eta_{ab})$, as in Special Relativity, and allowing for the more general structure of a 4-dimensional manifold with a (generally curved) *dynamic* metric $(\mathcal{M}, g_{ab})$, which is not *a priori* defined, but rather must be determined jointly with the matter evolving under its influence.

This dynamical content of GR may be very elegantly summarized through the Einstein Equations, which govern how the matter content in spacetime acts as a source for its curvature (for the definitions of curvature tensors and covariant derivatives, see Appendix B):

$$R_{ab} - \tfrac{1}{2} R g_{ab} = -8\pi G T_{ab}, \tag{3.1}$$

where T_{ab} is the matter stress tensor and we have kept Newton's constant G for later convenience in chapter 4, where we show that it can be renormalized as a coupling constant between matter and spacetime. Equation (3.1) is almost the most general covariant second-order equation which automatically leads to the covariant quantization of the stress tensor[A] that one can write for g_{ab}; the most general form is achieved by simply adding a term proportional to g_{ab}, introducing a cosmological constant Λ,

$$R_{ab} - \tfrac{1}{2} R g_{ab} + \Lambda g_{ab} = -8\pi G T_{ab}. \tag{3.2}$$

These equations, eventually supplemented by the equations of motion of matter, and initial/boundary conditions, will allow us to predict the geometry of all spacetime and of the matter propagating within it.

3.1.1 Lagrangian Formulation of General Relativity

Both in classical particle mechanics and field theories in flat spacetime, one can express their entire dynamical content through their equations of motion (such as Newton's or Maxwell's equations). It allows one to tell the evolution of a system from given initial

[A] As we have derived in appendix B, the Bianchi identity implies that the covariant derivative of the LHS of Einstein equations should be null (*i.e.*, $\nabla_a G^{ab} = 0$). Thus the same must be true for the RHS.

conditions and thus to make any possible physical predictions on it. Similarly, in General Relativity, its dynamical content can be fully expressed through Einstein's equations (3.1).

However, just as it happens with the former theories, it is desirable to present GR with a Lagrangian (or Hamiltonian) formulation for a number of reasons. Besides aesthetical and simplicity considerations, our known methods of quantization employ either of these formulations. Thus, not only do they prove central if one attemps to quantize gravity through a recognizable approach, but also they are necessary for QFTCS (and semiclassical gravity), so that one has a formulation of field theory in curved space liable to quantization (and is eventually able to connect quantum fields as a source of curvature for classical spacetimes).

For the purposes of this work, it will suffice for us to present only a Lagrangian Formulation[B]. It has the advantages of providing a manifesly covariant description of our theories (whereas a Hamiltonian relies on a split between space and time), and of allowing us to very simply obtain our dynamical equations. The mere existence of the correspondence with a Hamiltonian formulation will allow us to directly implement the scheme of canonical quantization, but rather than applying it to configuration and momenta variables, we impose the commutation relations directly to field mode operators.

With that said, we turn our attention to the construction of a Lagrangian formulation of GR. Before discussing a full dynamic theory of spacetime and matter, and elaborating on how we may adapt the latter to include gravity, let us begin by showing how we can encompass the spacetime geometry alone in a Lagrangian formulation, and obtain the vacuum Einstein Equations through an action principle.

Generally, for field theories, we have been considering an action functional S which only depends on its field variables *locally*, in the form of a spacetime integral of a scalar Lagrangian[C] function $\mathscr{L}$ (see eq. 2.13). Here, we want to build a purely geometrical action, which will likewise be constructed from a local scalar function of the metric, $S_G[g_{ab}]$, to be written in the form:

$$S_G = \int_{\mathcal{M}} d\mu_g(x) \mathscr{L}_G\Big(g_{ab}(x)\Big), \tag{3.3}$$

where $\mathscr{L}_g\Big(g_{ab}(x)\Big)$ depends only on the metric and its spacetime derivatives (which shall appear through curvature terms) at the event x, and $d\mu_g(x)$ is the natural volume element

[B] For a Hamiltonian formulation of GR, we refer the reader to Appendix E of (5), on which much of the presentation of the Lagrangian formulation in the present section is based.

[C] Technically, this is what we called a *Lagrangian density* in Chapter 2, where we reserved the term 'Lagrangian' to spatial integrals of $\mathscr{L}$. From this point onwards we shall refer to $\mathscr{L}$ only as the Lagrangian; the term *Lagrangian density* will be assigned with a different meaning below.

in the spacetime manifold $\mathcal{M}$[D] *induced by the metric* g_{ab}. Thus, compared with theories analyzed in the last chapter, GR presents us with a difficulty. If we attempt to look at variations of S_G with respect to the metric, we are faced with the awkward convolution that not only $\mathscr{L}_G$ but also the volume element itself depend on g_{ab}. To circumvent this, we begin by noting that the covariant (coordinate-independent) volume element can be expressed in any coordinate system as $d\mu_g(x) = d^4x|g(x)|^{1/2}$, being d^4x a (coordinate-dependent) coordinate volume element in $\mathbb{R}^4$ and $|g(x)|^{\frac{1}{2}}$ the Jacobian associated to it; its dependence on the metric can be simply codified as a determinant of its components in the coordinate basis[E](5, 19): $|g(x)| = |\det(g_{\mu\nu}(x))|$.

Now that we have properly isolated the metric dependence in the volume element, one particularly convenient way to handle it is to absorb this dependence in the integrand and perform the integrations in the (metric-independent) *coordinate volume*. To do so, we define *tensor densities* as follows: given any *tensor* field $T^{def...}_{abc...}$, whose definition does not make reference to any particular coordinate system, *we construct an associated tensor density field* $\tilde{T}^{def...}_{abc...}$ *in a given coordinate system* by defining its value in each point as: $\tilde{T}^{def...}_{abc...}(x) \equiv |g(x)|^{1/2} T^{def...}_{abc...}(x)$. Particularly, for a scalar field $\mathscr{L}$ we will have an associated *scalar density* $\tilde{\mathscr{L}} = |g|^{\frac{1}{2}}\mathscr{L}$.

With these considerations, let us show how the vacuum Einstein equations may be very elegantly obtained from what is arguably the most simple nontrivial action one can build from purely geometrical scalars. Postulating the Lagrangian $\mathscr{L} = R$, or equivalently, the *Lagrangian density* $\tilde{\mathscr{L}} = |g|^{1/2}R$, we obtain the famous Einstein-Hilbert action:

$$S_G[g^{ab}] = \int d^4x|g(x)|^{1/2}R(x), \tag{3.4}$$

where, for a matter of convenience, we are regarding S_G as a function of the *inverse metric* g^{ab}, rather than of g_{ab}. Now, using the same apparatus as in section 2.1, let us explicitly show how to compute its functional derivatives and obtain the associated dynamical equations. It proves convenient to evaluate them in the form of infinitesimal variations:

$$\delta(|g|^{1/2}g^{ab}R_{ab}) = \delta(|g|^{1/2})g^{ab}R_{ab} + |g|^{1/2}\delta g^{ab}\, R_{ab} + |g|^{1/2}g^{ab}\delta R_{ab}. \tag{3.5}$$

The second term is already proprortional to a variation in the argument g^{ab}. The

[D] For more details on integrations in manifolds, and volume elements see appendix B of (5). Further reference on the subject can be found in chapter I of (19).

[E] In fact, this was already true for flat spacetimes, but there are two key differences: (i) there, one can always find globally inertial coordinates, making Jacobian $|\eta(x)|^{\frac{1}{2}}$ trivially 1, and (ii) while there the metric was merely a background structure, here it is a dynamical variable and we must compute variations with respect to it.

first is also relatively straightforward to compute in terms of it, as it is a direct function of g^{ab}:

$$\delta|g|^{1/2} = -\tfrac{1}{2}|g|^{-1/2}\delta g = +\tfrac{1}{2}|g|^{1/2}\Big[g^{-1}\delta g\Big] = \tfrac{1}{2}|g|^{1/2}g^{ab}\delta g_{ab} = -\tfrac{1}{2}|g|^{1/2}g_{ab}\delta g^{ab}. \tag{3.6}$$

(Here we stress that, in the middle equality, we have rewritten a product involving the variations of the metric *determinant* in terms of the trace of product involving variations of the metric *tensor*.)

The third term, however, involves a variation in curvature. This makes it somewhat more convoluted to compute, since its relation to the metric is only indirectly defined through covariant derivatives. We make a more complete discussion on how to compute these variations in appendix B, from which we merely quote the result:

$$\delta R_{ab} = \tfrac{1}{2}g^{cd}[\nabla_a\nabla_b\delta g_{cd} + \nabla_c\nabla_d\delta g_{ab} - 2\nabla_c\nabla_{(b}\delta g_{a)d}]. \tag{3.7}$$

Thus, we have that the third term in (3.5) is proportional to:

$$g^{ab}\delta R_{ab} = \nabla^a\nabla_a(g^{cd}\delta g_{cd}) - \nabla^a\nabla^b\delta g_{ab} = \nabla^a v_a, \tag{3.8}$$

where we have defined: $v_a \equiv \nabla_a(g^{cd}\delta g_{cd}) - \nabla^b\delta g_{ab}$.

Thus, we see this term takes the form of a perfect divergence, making only a boundary contribution to the variations in S_g. Since boundary terms do not make any contribution to the local degrees of freedom in the bulk (and thus to the dynamic equations), we temporarily just ignore this term and obtain the following variation:

$$\frac{\delta S_G}{\delta g^{ab}} = |g|^{1/2}\Big(R_{ab} - \tfrac{1}{2}Rg_{ab}\Big). \tag{3.9}$$

Thus, extremizing the Einstein-Hilbert action,

$$\frac{\delta S_G}{\delta g^{ab}} = 0, \tag{3.10}$$

we obtain precisely Einstein's equations in the vacuum:

$$R_{ab} - \tfrac{1}{2}Rg_{ab} = 0. \tag{3.11}$$

Before we proceed, we briefly comment on the matter of boundary terms. Usually, such terms make no contribution to δS whatsoever, provided that one forces the variations of the relevant field (in our case δg^{ab}) to vanish at the boundary. *However, this is actually not the case for the Einstein-Hilbert action*; due to the fact that R involves second derivatives of the metric, one must *also require* that the derivativatives of the variations $\nabla_c \delta g_{ab}$ vanish at the boundary, or else define a boundary *counterterm* to subtract in the action. In the scope of this work, we shall not occupy ourselves with these boundary terms. The interested reader can find a few comments on the subject in the aforementioned Appendix E of (5), and a quite thorough discussion in (16).

At this point, one can very simply incorporate a cosmological constant Λ to the Einstein equations simply by adding a constant term Λ in the Einstein-Hilbert Lagrangian. More precisely, by making: $\tilde{\mathscr{L}}_G = |g|^{1/2}(R - 2\Lambda)$. The last term only yields a variation due to (3.6), whereupon one can easily verify that its addition brings us from (3.11) to:

$$R_{ab} - \tfrac{1}{2} R g_{ab} + \Lambda g_{ab} = 0. \tag{3.12}$$

At this point, we take the chance to note that Einstein equations (with or without a cosmological constant) are nontrivial in 4 (or more) dimensions. In this case, spacetime alone turns out to have local degrees of freedom, which counting the metric symmetries and all its nondynamical components related to gauge symmetries, the number of degrees of freedom amount to 2 per point in space, which will correspond to two independent polarizations of gravitational waves.

Now, we will show how we can incorporate matter in this formalism, and provide a full general-relativistic theory of matter and curved spacetime. Our previous requirements that the (matter) Lagrangian must be a scalar and that it takes a covariant form can be quite directly transported to curved spacetime through the prescription known as "minimal substitution". It goes as follows: given a special-covariant theory, defined in Minkowski spacetime by a Lagrangian involving the metric $\eta_{\mu\nu}$ and spacetime derivatives ∂_μ, one shall everywhere substitute $\eta_{\mu\nu} \to g_{\mu\nu}$ and $\partial_\mu \to \nabla_\mu$, making it generally covariant in curved spaces[F]. Thus, for instance, the Klein-Gordon field (2.29) minimally substituted in curved space would be:

$$\mathscr{L}_M = \frac{1}{2} g^{\mu\nu} (\nabla_\mu \phi)(\nabla_\nu \phi) - \frac{m^2}{2} \phi^2. \tag{3.13}$$

We stress that this procedure is by no means the only possible generalization of special-covariant theories to curved spacetimes (one could, for instance, add covariant

[F] For a more detailed discussion on special and general covariance, as well as the notions of covariance in prerelativistic physics, see chapter 4 of (5).

terms proportional to curvature, which will vanish as $g_{\mu\nu} \rightarrow \eta_{\mu\nu}$, recovering (2.29) in flat space), nor is it always free of ambiguities (as when one has 2 equivalent formulations in flat space, in terms of fields or of potentials, and these do not necessarily remain equivalent in curved space after minimal substitution (5)). Nevertheless, it is a consistent and practical prescription, and often the first one has at hand when trying to generalize a theory to curved spacetimes.

Now, as the geometrical portion of the action S_G does not depend on any matter fields, the dynamical equations of the latter (*i.e.* the Euler-Lagrange equations) will spring solely from the matter portion S_M:

$$S_M[\phi_a, g^{bc}] = \int_{\mathcal{M}} d^4x |g(x)|^{\frac{1}{2}} \mathscr{L}_M\Big(\phi_a, g^{bc}\Big). \tag{3.14}$$

Note that, although S_G does not carry any dependence on the matter fields ϕ_a, S_M necessarily depends on the metric. This codifies the fact that our fields are propagating through curved spacetimes, and will necessarily be influenced by its geometry. We obtain their Euler-Lagrange equations by extremizing S_M with respect to the fields:

$$\frac{\delta S_M}{\delta \phi_a} = 0. \tag{3.15}$$

For instance, for our minimally substituted scalar field (3.13):

$$g^{\mu\nu}\nabla_\mu\nabla_\nu\phi + m^2\phi \equiv \Big[\Box + m^2\Big]\phi = 0, \tag{3.16}$$

which turns out formally identical to (2.30), as we defined the general-covariant D'Alembertian $\Box \equiv g^{\mu\nu}\nabla_\mu\nabla_\nu$, although (2.30) and (3.16) are different (nonequivalent) equations!

On the other hand, to look at the effect that matter has on spacetime, acting as a source of curvature, we must consider the entire action:

$$S[\phi_a, g^{ab}] = S_G[g^{ab}] + S_M[g^{ab}, \phi_a]. \tag{3.17}$$

When we first defined S_G, we were not worried about its normalization, as it turned out superfluous for the vacuum equations. To reobtain Einstein's equations with a source, however, one must adjust a relative normalization between S_G and S_M (and, of course, one must assure both terms dimensionally consistent, although this matter is entirely hidden in Planck Units, and partially hidden in natural units). This can be achieved by readjusting $\mathscr{L}_G$ as:

$$\mathscr{L}_G \equiv \frac{R - 2\Lambda}{16\pi G}. \tag{3.18}$$

Then, by extremizing the total action with respect to the metric,

$$\frac{\delta S}{\delta g^{\mu\nu}} = \frac{|g|^{\frac{1}{2}}}{16\pi G}\Big(R_{\mu\nu} - \tfrac{1}{2} R g_{\mu\nu} + \Lambda g_{\mu\nu}\Big) + \frac{\delta S_M}{\delta g^{\mu\nu}} = 0, \tag{3.19}$$

we obtain a natural definition for the stress tensor of the matter fields, so that we recover the full Einstein equation (3.2):

$$T_{ab} \equiv \frac{2}{|g|^{\frac{1}{2}}} \frac{\delta S_M}{\delta g^{ab}}. \tag{3.20}$$

Finally, we note that, the imposition that both S_G and S_M must be scalars, and thus invariant under any spacetime transformations (or, equivalently, any coordinate transformations) will imply that G_{ab} and T_{ab} must be both covariantly conserved, regardless of the Einstein equations.

In direct analogy to what we did in the past chapter, it would seem like a very natural next step to try to quantize this full theory of gravity and matter (or perhaps gravity alone, for a start), imposing some procedure of quantization to the fields ϕ_a and g_{ab}. However, the attempts to carry out a quantization for spacetime itself, be it through a metric field or more profound changes in the whole spacetime structure classically described by $(\mathcal{M}, g_{ab})$, have met enormous challenges in the past decades, so that we are still far from a satisfactory solution for such a theory[G]. A far more manageable approach is to quantize matter fields alone in a classical curved background geometry, which gives rise to what we call *quantum field theory in curved spaces*. This approach consists of finding a way to generalize some of the procedures and basic definitions originally carried in flat, Minkowski background space to generally curved background spaces; it has proved quite successful in describing phenomena for quantum matter fields in which curvature plays a relevant role, but is not too extreme so that it itself does not need to be considered quantized. Furthermore, this curved space theory has raised many relevant questions to QFT in Minkowski spaces, which were previously unnoticed due to the fact that its traditional approaches relied heavily on the Poincaré group of symmetries.

Then, it is this approach of quantum field theory in curved space that we shall develop throughout the rest of this chapter. Before we can proceed to it, however, we shall

[G] Again, for a thorough review on the state-of-the-art of many contemporary approaches to quantum gravity, see (15) and references therein.

take a moment in the next section to impose appropriate restrictions in our curved spaces, so that we can make meaningful extensions of many of the concepts defined in Minkowski space and use them for quantization in curved ones.

3.1.2 Spacetime Geometry and Quantum Field Theory

We have just seen how General Relativity provides a quite natural framework to analyze the mutual dynamic of matter and spacetime in a classical context. Indeed, it gives us well-defined local dynamical equations, which should allow us to predict the behaviour of matter and geometry from an appropriate set of initial conditions[H]. However, the curved nature of spacetime in GR confers it with a few subtleties and complications for the initial value formulation, when compared to flat space. Although classical field theories are not indifferent to these subtleties – particularly in terms of predictability and a well-posed initial value formulation–, they manifest more acutely in the quantum case, where nonlocal features play a more proeminent role in theory and, particularly, a notion of 'equal-time' surfaces is required to postulate the canonical commutation relations in the Hamiltonian formalism. With those matters in mind, we give a brief account of the necessary structure of spacetime to our present formulation of QFTCS, with particular emphasis on its causal structure. This is intended to be just an overview on the topic, sufficient to situate the unfamiliar reader in the subsequential discussion; for a more complete account of the subject, we refer the reader to (5, 10, 11, 25)[I], which are the direct sources of the present exposition.

In a pregravitational context, thoroughly discussed in the last chapter, we have seen that a crucial structure to the initial value formulation (*i.e.* to obtain a unique solution from the field equations with a given initial condition), as well as to the field mode decomposition and to the postulation of canonical commutation relation, was *equal-time surfaces*. These surfaces allowed us to speak meaningfully of field configurations ('at a given time') and perform "complete" spatial integrations, for example for the Poisson brackets (2.27) and the inner product (2.39). There, since we were handling either Galilean or Minkowskian spacetimes, where we have either absolute time or a very simple notion of flat equal-time surfaces attached to congruences of inertial worldlines (*i.e.*, to families of inertial observers), we have restricted our analysis to these simple surfaces.

In GR [J], the curved nature of spacetime does not generally allow for such a distinct and simple construction of 'equal-time' surfaces. Notwithstanding, we shall see that for a

[H] This is generally known as the Initial Value Problem (IVP), or Boundary and Initial Value Problem (BIVP), when spatial boundary conditions are also required. In-depth discussions of the IVP in GR can be found in chapter 10 of (5) and in chapter 7 of (25).

[I] We warn, however, that they require basic notions of topology for a fluid reading.

[J] Or in any modified-gravity theories that share the basic spacetime structure as a manifold with a pseudo-Riemannian metric $(\mathcal{M}, g_{ab})$.

quite general class of spacetimes, the so-called *globally hyperbolic* spacetimes, one has a generalized notion of simultaneity surfaces, whose causal domains extend to the entire spacetime: *Cauchy Surfaces*. In order to properly define the latter, and provide a little physical intuition on them, we go over a few basic concepts on the causal structure of spacetime.

It follows immediately from the equivalence principle – which states that, *locally*, any curved spacetime 'looks like' flat (Minkowski) spacetime: that is, one can always construct a *local inertial frame* such that the metric components $g_{\mu\nu}$ at an event x are equal to $\eta_{\mu\nu} = (+1; -1, -1, -1)$ and its first derivatives vanish – that the *local* causal structure of general-relativistic spacetimes is the same as in Minkowski spacetime. Its *global* structure, however, may differ radically, for instance due to nontrivial topologies, to the "tipping of light cones", or even to singularities. Let us then classify causal structures and point out some desirable features for a 'well-behaved' spacetime (our counterexamples may seem particularly artificial at times, but that is in part recourse to pedagogical examples).

- Time Orientability: A very basic property we would like for physically plausible spacetimes is the possibility to determine, for every event x, its past and future directions, and unambiguosly distiguish them. Locally, this is done by constructing the light cones around each event, which divides all events with a positive, timelike separation in two disconnected regions ("above" and "below" the light cone); one can then identify one of them with the (chronological) past and the other with the (chronological) future of that event. Trouble may arise, however, when we try to extend this identification globally. In Minkowski, this can be trivially achievable through the affine structure of space: if one chooses a fiducial event, traces its light cone, and identifies its past and future, one needs simply to translate this rigidly through all spacetime to obtain a unique and consistent identification; equivalently, one may identify future-directed (past-directed) timelike vectors in any two events directly[K].

In curved spacetimes, however, one cannot automatically identify the tangent vectors in distinct events. The best one can do is to identify them continuously through parallel transport. However, due to spacetime curvature, there will generally be a "tipping" of the light cones throughout spacetime. In extreme cases, this tipping may result in a loss of global orientability of the space (something like in Moebius strip kind of spacetime), such that, along a closed curve one may "tip the light cone upside-down" and be unable to obtain a globally consistent time orientation (see figure 13).

Then, the first and most basic requirement that we shall make for our spacetimes is that they are time-orientable.

-Chronal/Causal past and future: For any time-orientable spacetime $(\mathcal{M}, g_{ab})$,

[K] A timelike vector is said to be future-directed (past-directed) if it is in the future (past) section of the interior of the light cone.

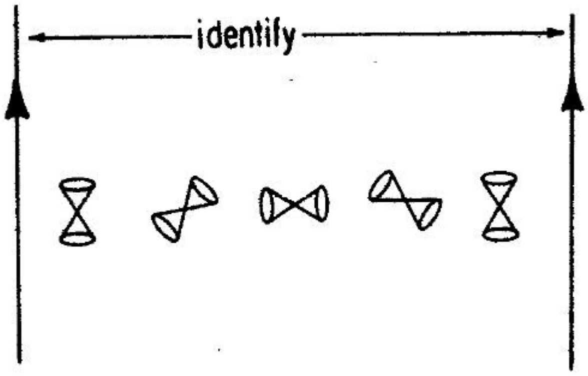

Figure 13 – Pictoric example of a non time-orientable manifold. Here the light cones tipp over 180 degrees as one trancurs a full period from left to right, making it impossible to consistently identify past and future directions.
Source: WALD (5)

one can identify the *chronological future* of each event x, $I^+(x)$, the set of all events that can be reached by a future-oriented timelike geodesic[L] at a strictly positive proper-time interval $|\Delta\tau| > 0$ (we require that $|\Delta\tau| \neq 0$ to leave out a curve of null arclength taking x into x). Likewise, we define its *cronological past*, $I^-(x)$, as all events that can be reached by a past-directed timelike geodesic at a strictly positive proper-time interval. Then, for any two events $x, y \in \mathcal{M}$, it is obvious that $y \in I^+(x) \Leftrightarrow x \in I^-(y)$. Similarly, we define the *causal future* (*causal past*) of x, $J^+(x)$ ($J^-(x)$) as the set of all events that can be reached by future-directed (past-directed) timelike *or null* geodesics (these are collectively called *causal geodesics*). Note that, unlike its chronological future (past), this encompasses the possibility of null length curves, so that we always have that $x \in J^+(x)$ and $x \in J^-(x)$.

Now, one can see that if in a spacetime $(\mathcal{M}, g_{ab})$, there are events such that $x \in I^+(x)$, this will mean that $\mathcal{M}$ possesses nontrivial closed timelike curves. A quite straightforward example of a spacetime that does possess closed timelike curves is a flat "timelike-torus", which can be obtained from Minkowski spacetime by identifying two equal-time surfaces $t = 0$ and $t_0 > 0$[M]. Such spaces are generally regarded as unphysical, and may lead to paradoxes as events may lie in their own chronological future. Thus, generally, we shall require that the spacetimes we are considering do not possess any closed timelike curves, such that $x \notin I^+(x)$, $\forall x \in \mathcal{M}$.

For any subset $A \subset \mathcal{M}$ we can define its chronal and causal pasts and futures, $I^\pm(A)$, $J^\pm(A)$ as the union of the respective regions for each of their events, that is:

[L] A geodesic $\gamma(t)$ is said to be future-directed as a function of the parameter t if its tangent vector $(\frac{\partial}{\partial t})^a$ is everywhere future-directed.

[M] One may argue that such a spacetime is too "artificial", being produced merely by strange topological identifications. However, one can more generally build solutions with closed timelike curves without such topological identifications, such as Gödel's Universe (see section 7.7 of (25)).

$$I^{\pm}(A) \equiv \bigcup_{x \in A} I^{\pm}(x), \qquad J^{\pm}(A) \equiv \bigcup_{x \in A} J^{\pm}(x). \tag{3.21}$$

-Achronal sets: An important definition to start to encompass the notion of an equal time surface is that of achronal sets. A subset $S \subset \mathcal{M}$ is said to be achronal if no two events x and y belonging to it are chronologically related (*i.e.*, if $y \notin I^{\pm}(x)$, $\forall x, y \in S$), that is:

$$I^{+}(S) \cap S = \emptyset. \tag{3.22}$$

This definition prevents one from obtaining an inconsistent notion of simultaneity, as any useful notion of simultaneous events will certainly exclude events that are in the chronological past or future of one another (note that this will only be possible throughout spacetime if it does not have closed timelike curves, such that no event can lie in its own causal past/future).

A particular class of achronal sets that will be of interest to us is that of spacelike differentiable surfaces in $\mathcal{M}$. Although these generally allow one to identify a notion of 'simultaneity' in spacetime, they still do not emcompass all we need for a well-posed initial value formulation. For that, we must still require that they are, in a sense, complete. We shall give that a more precise meaning through the definition of domains of dependence.

-Domains of dependence and Cauchy Surfaces: given any achronal set S, we define its future domain of dependence $D^{+}(S)$ as the set of all events y, for which *any* past-inextendible[N] causal geodesics intersecting y will intersect S. Likewise, one defines its past domain of dependence, $D^{-}(S)$ as the set of all events x, for which *any* future-inextendible causal geodesics intersecting x will intersect S. One then defines its total domain of dependence $D(S)$ as the union $D^{+}(s) \cup D^{-}(s)$.

This notion is very important for the initial formulation of any causal field theory because, if information from a field can only be transported along causal curves, then knowledge from the field (and its independent derivatives) at an achronal surface S will allow us to determine the field throughout all $D(S)$ (see Figure 14).

[N] A causal curve γ in $\mathcal{M}$ is said to be past inextendible if it has no past endpoints. That means it will either run off to infinity or 'fall in an edge' of spacetime (such as singularity). One can similarly define future-inextendible curves.

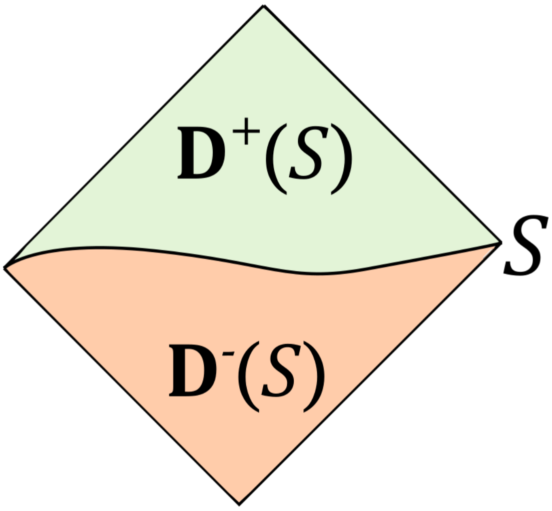

Figure 14 – Domains of dependence of a compact achronal set S (represented as a curvy line in the middle). Its future domain of dependence, $D^{+}(S)$, is represented in green and its past domain of dependence $D^{-}(S)$ is represented in orange. In the 4-dimensional compact region formed by $D(S)$, one can predict the configurations of a matter field solely from knowledge from it on S.
Source: By the author.

Finally, we are at a place to define a useful generalization of equal-time surfaces in curved spaces, in terms of which we can have a well-posed initial value formulation: *Cauchy surfaces*. A closed achronal surface Σ is said to be a Cauchy surface if its domain of dependence extends to the entire spacetime, that is: $D(\Sigma) = \mathcal{M}$. Spacetimes which possess Cauchy surfaces are called *globally hyperbolic*. The most important property of globally hyperbolic spacetimes is that they will allow us to predict the state of a field at any event if we have completely determined its state (configurations and derivatives) at a given 'instant of time', that is, at a given Cauchy surface (see Figure 15).

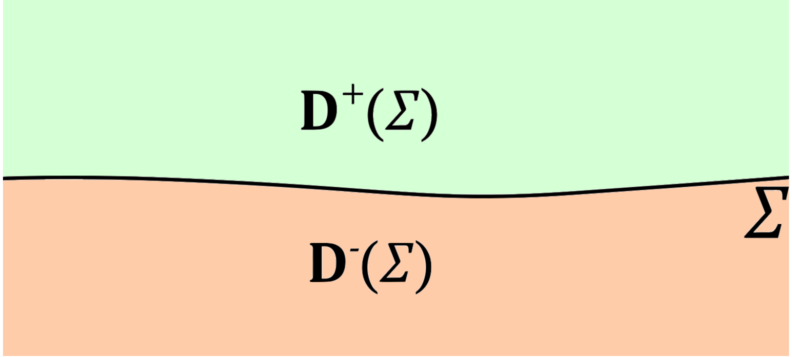

Figure 15 – A Cauchy surface Σ and its domains of dependence, $D^{+}(S)$ (green) and $D^{-}(S)$ (orange). In this case we see that $D(\Sigma) = \mathcal{M}$, such that one can predict dynamical information in the entire spacetime if one has appropriate initial conditions at Σ. (Here, one should imagine Σ and $D^{\pm}(\Sigma)$ as extending all the way to infinity.)
Source: By the author.

Then, in the following sections, we will require our background spacetimes to be always smooth (so that we may define derivatives to any order) and globally hyperbolic pseudo-Riemannian manifolds; these will be sufficient conditions for us to define a well-posed quantized theory in them.

3.2 Quantization of a Scalar Field

With these classical foundations at hand, we are in position to extend the formalism of chapter 2 and carry out the quantization of noninteracting fields *in a curved background spacetime* $(\mathcal{M}, g_{ab})$ – *i.e.* in a classical spacetime with a *given* curved metric g_{ab}. We shall implement this procedure by appealing to the existence of complete sets of normal modes $\{u_i, u_i^*\}$ to our linear field equations, and promoting the classical amplitudes α_i of these modes to linear operators a_i (in a suitable Hilbert Space $\mathcal{H}$), upon which we impose the mode commutation relations analogous to (2.62). In summary:

- Classical field: $\phi(x) = \sum_i \alpha_i u_i(x) + \alpha_i^* u_i^*(x)$
- $\alpha_i, \alpha_i^* \in \mathbb{C} \longrightarrow a_i, a_i^\dagger \in GL(\mathcal{H}), \qquad [a_i, a_j^\dagger] = \delta_{ij}, \quad [a_i, a_j] = 0 = [a_i^\dagger, a_j^\dagger]$
- Quantized field: $\phi(x) = \sum_i a_i u_i(x) + a_i^\dagger u_i^*(x)$

With this basic prescription in mind, let us develop such a process more explicitly. As in the last chapter, we take a real scalar field as a working model. Consider a field $\phi(x)$ with a Lagrangian density:

$$\tilde{\mathscr{L}} = \sqrt{-g}\Big(\tfrac{1}{2}g^{ab}(\nabla_a\phi)(\nabla_b\phi) - \tfrac{1}{2}[m^2 + \xi R]\phi^2\Big). \tag{3.23}$$

This has the usual kinetic and mass terms, obtained directly from the flat space theory via "minimal substitution" (see (3.13)), as well as a local 'nonminimal' covariant coupling with the spacetime curvature, expressed in the term $\frac{1}{2}\xi R\phi^2$. Special interest attaches to the values $\xi = 0$ (minimal coupling) and $\xi = 1/6$ (conformal coupling, in 4 spacetime dimensions; see appendix B). From the Lagrangian density (3.23), we immediately obtain the action and derive the dynamic equations for ϕ:

$$S = \int d^4x\, \tilde{\mathscr{L}}(x) = \int d^4x\, \sqrt{-g}\Big(\tfrac{1}{2}g^{ab}(\nabla_a\phi)(\nabla_b\phi) - \tfrac{1}{2}[m^2 + \xi R]\phi^2\Big), \tag{3.24}$$

$$\frac{\delta S}{\delta\phi(x)} = 0 \quad \Rightarrow \quad \Big[\Box_x + m^2 + \xi R(x)\Big]\phi(x) = 0, \tag{3.25}$$

where we have defined the D'Alembertian operator *in curved spacetime* as $\Box = g^{ab}\nabla_a\nabla_b$. When acting on a scalar field, it may be written in terms of mere partial derivatives in the form $\Box\phi = |g|^{-1/2}\partial_\mu(|g|^{1/2}g^{\mu\nu}\partial_\nu\phi)$ (see eq. (B.8)).

Since we are working with a fixed background geometry – that is, we are ignoring the gravitational effects of ϕ in the metric –, eq. (3.25) will indeed be a linear second-order PDE, such that any of its solutions can be expanded in a given basis of modes.

Similarly to the case of flat spacetime, we define an inner product which will allow us to compute projections and decompose any solutions in a given set of modes. These projections will allow us to extract the maximal information of our field from some set of initial conditions in a Cauchy surface Σ. We make our generalization as follows: we foliate our (globally hyperbolic) spacetime by an arbitrary family of Cauchy surfaces, and we pick *any* surface Σ from this family to compute:

$$\begin{aligned}(\phi,\psi) &\equiv i\int_\Sigma d^3x|g_\Sigma(x)|^{\frac{1}{2}}\, n^\mu(x)\phi^*(x)\overleftrightarrow{\partial_\mu}\psi(x)\\ &= i\int_\Sigma d^3x|g_\Sigma(x)|^{\frac{1}{2}}\, n^\mu(x)\Big(\phi^*(x)\partial_\mu\psi(x) - (\partial_\mu\phi^*(x))\psi(x)\Big),\end{aligned} \tag{3.26}$$

where $n^\mu(x)$ is the unitary, future-directed vector normal to Σ at $x \in \Sigma$, and d^3x the coordinate 3-volume element. $(-g_\Sigma)_{ij}$ is the (positive-definite) metric induced on Σ by $g_{\mu\nu}$, so that the (coordinate independent) induced volume element on Σ is $d\mu_{g_\Sigma}(x) = |g_\Sigma(x)|^{\frac{1}{2}}d^3x$.

Just as with (2.39), the definition above allows us to immediately verify the elementary properties (2.40) of sesquilinearity. In spite of these useful properties, since our definition relies on an arbitrary choice of integration surface (which, furthermore, is not related to any 'special' family of observers), it may not be *a priori* obvious that this product bears similar physical significance to (2.39), and whether arbitrary (surface-dependent) elements might appear in it. Indeed, as it happened in flat space, for a completely arbitrary pair of scalar functions u, w in $\mathcal{M}$, the result of (3.26) will obviously depend on the choice of Σ. We shall show, however, that if these functions are solutions of the field equations (3.25), then their scalar product is independent of Σ.

The proof is very similar as in flat space: again, we consider the difference of the product evaluated in two surfaces Σ and $\Sigma' \subset I^+(\Sigma)$ and write them as a boundary term; then, using Gauss's theorem, we express it as a volume integral, which will be identically vanishing for any two functions obeying the field equations (see Figure 16):

$$\begin{aligned}(u,w)_{\Sigma'} - (u,w)_\Sigma &= \int_{\Sigma'} d\mu_{g_\Sigma}(x)\; n^\mu(x)\, u^*(x)\overleftrightarrow{\partial}_\mu w(x) - \int_\Sigma d\mu_{g_{\Sigma'}}(x)\; n^\mu(x)\, u^*(x)\overleftrightarrow{\partial}_\mu w(x)\\ &= \int_v d\mu_g(x)\; \nabla^\mu(u^*(x)\overleftrightarrow{\partial}_\mu w(x))\end{aligned}$$

$$
\begin{aligned}
&= \int_v d\mu_g(x)(u^*(x)\nabla^\mu\nabla_\mu w(x) - w(x)\nabla^\mu\nabla_\mu u^*(x)) \\
&= 0. \qquad (3.27)
\end{aligned}
$$

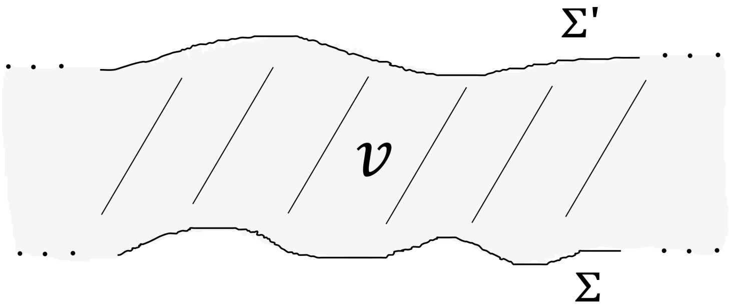

Figure 16 – Spacetime volume v (grey hatched area) whose bondary is composed by the two Cauchy surfaces Σ and $\Sigma' \subset I^+(\Sigma)$.
Source: By the author

Armed with this inner product, it is possible to find an orthonormal basis of solutions to the field equations $\{u_i(x), u_i^*(x)\}$:

$$
(u_i, u_j) = \delta_{ij} = -(u_i^*, u_j^*), \qquad (u_i, u_j^*) = 0, \qquad (3.28)
$$

so that we can then expand the classical field in the form:

$$
\phi(x) = \sum_i \alpha_i u_i(x) + \alpha_i^* u_i^*(x)\,. \qquad (3.29)
$$

Now, we can proceed to quantization in an entirely analogous manner to (2.62), by promoting the classical mode amplitudes α_i, α_i^* to quantum operators $a_i, a_i^\dagger$ with the usual commutation relations[O]:

$$
[a_i, a_j] = [a_i^\dagger, a_j^\dagger] = 0, \qquad (3.30a)
$$

$$
[a_i, a_j^\dagger] = \delta_{ij}, \qquad (3.30b)
$$

such that the *quantized field operator* reads:

$$
\phi(x) = \sum_i a_i u_i(x) + a_i^\dagger u_i^*(x)\,. \qquad (3.31)
$$

[O] We stress that these will be *equivalent* to the canonical commutation relations, which can be defined for a given choice of foliation Σ_t in $\mathcal{M}$.

It then follows, just as in the case of flat spacetime, that we can build a C.S.C.O. in the Hilbert space of our theory from the number operators $N_i = a_i^\dagger a_i$ from an infinite collection of decoupled harmonic oscillators, and so define a Fock Space as usual. We define the vacuum state $|0\rangle$ as the one which is annihilated by all destruction operators a_i:

$$a_i \left|0\right\rangle = 0, \qquad \forall i. \tag{3.32}$$

And we can construct the n-particle states through successive applications of the creation operators $a_i^\dagger$:

$$\left|n_1, n_2 ...\right\rangle = \frac{1}{\sqrt{n_1! n_2! ...}} (a_1^\dagger)^{n_1} (a_2^\dagger)^{n_2} ... \left|0\right\rangle . \tag{3.33}$$

However, unlike in Minkowski spacetime, there are in general no "natural" sets of modes $\{u_i, u_i^*\}$ in terms of which to define a vacuum state. More precisely, there is no natural way to divide the space of solutions of the field equations in positive- and negative-frequency subspaces (spanned by modes $\{u_i\}$ and $\{u_i^*\}$, respectively). In the Minkowski case, we had a natural choice of coordinates (namely, globally inertial coordinates) and family of modes (plane waves) given by the Poincaré group of all isometries of the Minkowski spacetime. In particular t^a is a Killing field generating time translations, of which the plane waves $u_{\mathbf{k}}$ are eigenfunctions:

$$\begin{cases} i\partial_t u_{\mathbf{k}} = +\omega_k u_{\mathbf{k}}, & \text{(3.34a)} \\ i\partial_t u_{\mathbf{k}}^* = -\omega_k u_{\mathbf{k}}^*, & \text{(3.34b)} \end{cases}$$

where $\omega_k > 0$, $\forall \mathbf{k}$.

In a general curved spacetime, with no such symmetry to distinguish particular sets of modes, we could on equal footing consider a distinct set of normal modes, $\{\bar{u}_i, \bar{u}_i^*\}$, obeying the same orthonormality conditions (3.28), and write the classical field expansion as:

$$\phi(x) = \sum_i \bar{\alpha}_i \bar{u}_i(x) + \bar{\alpha}_i^* \bar{u}_i^*(x). \tag{3.35}$$

Then we could quantize ϕ by promoting these new mode amplitudes $\bar{\alpha}_i, \bar{\alpha}_i^*$ to operators $\bar{a}_i, \bar{a}_i^\dagger$ obeying the same commutation relations as (3.30):

$$\phi(x) = \sum_i \bar{a}_i \bar{u}_i(x) + \bar{a}_i^\dagger \bar{u}_i^*(x), \tag{3.36}$$

$$[\bar{a}_i, \bar{a}_j] = [\bar{a}_i^\dagger, \bar{a}_j^\dagger] = 0, \tag{3.37a}$$

$$[\bar{a}_i, \bar{a}_j^\dagger] = \delta_{ij}. \tag{3.37b}$$

Just as in the previous case, the number operators $\bar{N}_i = \bar{a}_i^\dagger \bar{a}_i$ make a C.S.C.O. in our Hilbert space, and we may once again define a Fock Space as usual, starting from a vaccum state $|\bar{0}\rangle$, annihilated by all $\bar{a}_i$:

$$\bar{a}_i \, |\bar{0}\rangle = 0, \qquad \forall i, \tag{3.38}$$

and similarly defining all particle states through successive applications of $\bar{a}_i^\dagger$ upon it.

Comparing the field expansions (3.31) and (3.36), each associated to their respective commutation relations and Fock Spaces, a few relevant questions arise. First of all, are both quantization procedures necessarily equivalent? (For instance, do they lead to mutually consistent field operator commutators? Are their Fock Spaces equivalent, and their physical predictons the same?) Second, do they share a common notion of vacuum? (That is, are $|0\rangle$ and $|\bar{0}\rangle$ always the same, perhaps up to a phase factor?) These questions turn out to reveal deep and interesting features of QFTCS, such as the nature (and inherent ambiguity) of the concept of particles, as well as surprising perspective on the connection between spin and statistics.

In order to address them, we must first specify the relations between these two sets of modes. Being both sets complete, they can each be expanded in terms of one another. For example, we could write $\bar{u}_i$ as:

$$\bar{u}_i = \sum_j (u_j, \bar{u}_i) u_j - (u_j^*, \bar{u}_i) u_j^* \; = \sum_j \alpha_{ij} u_j + \beta_{ij} u_j^* \,, \tag{3.39}$$

where the minus sign on the second term comes from the normalization (3.28). The coefficients α_{ij} and β_{ij} are known as *Bogolubov coefficients*, and are defined above as the projections:

$$\begin{cases} \alpha_{ij} \equiv (u_j, \bar{u}_i) = (\bar{u}_i, u_j)^* & (3.40\text{a}) \\ \beta_{ij} \equiv -(u_j^*, \bar{u}_i) = -(\bar{u}_i, u_j^*)^* & (3.40\text{b}) \end{cases} .$$

(Here, we warn that the exact conventions may vary somewhat in the literature.)

Conversely, we could expand the u_i modes in terms of $\bar{u}_j$ modes, as well as write similar expansions for the creation and annihilation operators. The latter expansions may be easily computed from the former (or vice-versa) by using the relations $a_i = (u_i, \phi)$ (or

$u_i = [\phi, a_i^\dagger]$). We summarize all these expansions in terms of the Bogolubov coefficients (3.40):

$$u_i = \sum_j \alpha_{ji}^* \bar{u}_j - \beta_{ji} \bar{u}_j^* \tag{3.41a}$$

$$\bar{u}_i = \sum_j \alpha_{ij} u_j + \beta_{ij} u_j^* \tag{3.41b}$$

$$a_i = \sum_j \alpha_{ji} \bar{a}_j + \beta_{ji}^* \bar{a}_j^\dagger \tag{3.42a}$$

$$\bar{a}_i = \sum_j \alpha_{ij}^* a_j - \beta_{ij}^* a_j^\dagger \tag{3.42b}$$

Also, the orthonormality and completeness of both sets of modes will imply in self-consistency properties for the Bogolubov coefficients. It is easy to deduce them by performing a back and forth transformation for any fixed mode $\bar{u}_i$:

$$\begin{aligned} \bar{u}_i &= \sum_j \alpha_{ij} u_j + \beta_{ij} u_j^* \\ &= \sum_k \left\{ \left[\sum_j \alpha_{ij} \alpha_{kj}^* - \beta_{ij} \beta_{kj}^* \right] \bar{u}_k - \left[\sum_j \alpha_{ij} \beta_{kj} - \beta_{ij} \alpha_{kj} \right] \bar{u}_k^* \right\}, \end{aligned} \tag{3.43}$$

which immediately implies:

$$\begin{cases} \displaystyle\sum_j \alpha_{ij} \alpha_{kj}^* - \beta_{ij} \beta_{kj}^* = \delta_{ik} & (3.44a) \\ \displaystyle\sum_j \alpha_{ij} \beta_{kj} - \beta_{ij} \alpha_{kj} = 0 & (3.44b) \end{cases} .$$

Let us then begin to investigate our first question. With the above relations and properties at hand, it is straightforward to verify whether the commutation relations (3.30) and (3.37) are mutually consistent. If we analize the covariant commutator $[\phi(x), \phi(y)]$ in both expansions, such consistency must imply the equality:

$$\sum_i u_i(x) u_i^*(y) - u_i(y) u_i^*(x) = [\phi(x), \phi(y)] = \sum_i \bar{u}_i(x) \bar{u}_i^*(y) - \bar{u}_i(y) \bar{u}_i^*(x). \tag{3.45}$$

Expanding the RHS of (3.45) by means of (3.41), we then obtain:

$$\begin{aligned} &\sum_i \Big\{ \Big(\textstyle\sum_j \alpha_{ij} u_j(x) + \beta_{ij} u_j^*(x) \Big) \Big(\textstyle\sum_k \alpha_{ik}^* u_k^*(y) + \beta_{ik}^* u_k(y) \Big) \\ &\qquad - \Big(\textstyle\sum_j \alpha_{ij} u_j(y) + \beta_{ij} u_j^*(y) \Big) \Big(\textstyle\sum_k \alpha_{ik}^* u_k^*(x) + \beta_{ik}^* u_k(x) \Big) \Big\} \\ = &\sum_{j,k} \Big[\Big(\textstyle\sum_i \alpha_{ij} \alpha_{ik}^* - \beta_{ik} \beta_{ij}^* \Big) u_j(x) u_k^*(y) - \Big(\textstyle\sum_i \alpha_{ik} \alpha_{ij}^* - \beta_{ij} \beta_{ik}^* \Big) u_k(y) u_j^*(x) \\ &\qquad + \Big(\textstyle\sum_i \alpha_{ij} \beta_{ik}^* - \alpha_{ik} \beta_{ij}^* \Big) u_j(x) u_k(y) + \Big(\textstyle\sum_i \beta_{ij} \alpha_{ik}^* - \beta_{ik} \alpha_{ij}^* \Big) u_j^*(x) u_k^*(y) \Big]. \end{aligned} \tag{3.46}$$

Then, comparing this to the LHS of (3.45), we get the following conditions for the Bogolubov coefficients:

$$\sum_i \alpha_{ij}\alpha^*_{ik} - \beta_{ik}\beta^*_{ij} = \delta_{jk}, \tag{3.47a}$$

$$\sum_i \alpha_{ij}\beta^*_{ik} - \alpha_{ik}\beta^*_{ij} = 0. \tag{3.47b}$$

But these conditions merely express the orthonormality and completeness of both bases (being equivalent to (3.44)), thus showing that (3.30) and (3.37) are indeed equivalent.

On the other hand, had we tried to quantize our fields by imposing anticommutation relations to these operators:

$$\{a_i, a_j\} = \{a^\dagger_i, a^\dagger_j\} = 0 = \{\bar{a}_i, \bar{a}_j\} = \{\bar{a}^\dagger_i, \bar{a}^\dagger_j\}, \tag{3.48a}$$

$$\{a_i, a^\dagger_j\} = \delta_{ij} = \{\bar{a}_i, \bar{a}^\dagger_j\}, \tag{3.48b}$$

we immediately see that the corresponding consistency check would yield:

$$\sum_i u_i(x)u^*_i(y) + u_i(y)u^*_i(x) = \{\phi(x), \phi(y)\} = \sum_i \bar{u}_i(x)\bar{u}^*_i(y) + \bar{u}_i(y)\bar{u}^*_i(x), \tag{3.49}$$

such that the corresponding conditions for the Bogolubov coefficients, namely:

$$\sum_i \alpha_{ij}\alpha^*_{ik} + \beta_{ik}\beta^*_{ij} = \delta_{jk}, \tag{3.50a}$$

$$\sum_i \alpha_{ij}\beta^*_{ik} + \alpha_{ik}\beta^*_{ij} = 0, \tag{3.50b}$$

are generally not satisfied for $\beta_{ij} \neq 0$. (In section 3.4, we shall reinterpret this result in terms of particle creation.)

Thus, we see that we can achieve consistency for quantization in two arbitrary families of modes (for a free scalar field) *only if we impose commutation relations* (rather than anticommutation relations) for our field mode operators. These, on their turn, will imply that ϕ must obey Bose-Einstein statistics.

Then, having assured mutual consistency between the commutation relations defined for any two bases of orthonormal modes, we would like to know the relation between their respective Fock spaces, particularly, how one may relate states $|n_1, n_2 \ldots\rangle$ defined with the occupation numbers of the modes $\{u_i\}$ in terms of $|\bar{n}_1, \bar{n}_2, \ldots\rangle$, defined with the occupation numbers of the modes $\{\bar{u}_i\}$.

To draw these relations, all one needs to determine is the general form of the projections $\langle \bar{n}_1, \bar{n}_2 ... | n_1, n_2 ... \rangle$ ($\propto \langle \bar{0} | 0 \rangle$). We shall not deduce the general form of those projections here[P], but we note that they can be computed with some algebraic effort employing the expansions (3.42) for the creation and annihilation operators (explicit expressions for them in terms of the Bogolubov coefficients can be found in (1), eqs. (3.45)-(3.47)). However, there are two important features of these amplitudes that we would like to point out: (i) the vacuum to many-particles transitions are only nonzero when the number of particles is even, as the creation (annihilation) operators $\bar{a}_i$ ($\bar{a}_i^\dagger$) are always linear in the operators $\bar{a}_i$ and $\bar{a}_i^\dagger$ and one needs an even number of such operators to match the number of created and annihilated particles and produce nonorthogonal states; (ii) these amplitude transitions are *always* proportional to the vacuum to vacuum amplitudes $\langle \bar{0} | 0 \rangle$, and thus it is necessary that $\langle \bar{0} | 0 \rangle \neq 0$ for both Fock spaces to represent the same Hilbert space.

When $\langle \bar{0} | 0 \rangle = 0$, the modes $\{u_i\}$ and $\{\bar{u}_i\}$ are said to yield unitarily inequivalent quantized theories. In the scope of the present work, we shall not go into detail of unitary (in)equivalence. For an in-depth discussion of the topic, see (8); also, the reader may find a quite simple example of unitarily inequivalent mode choices in Minkowski spacetime in the section II of (37), where one considers modes of the form:

$$\bar{u}_{\mathbf{k}} \propto (\alpha(k) e^{-i\omega_k t} + \beta(k) e^{i\omega_k t}) e^{i\mathbf{k} \cdot \mathbf{x}}. \tag{3.51}$$

Then, restricting ourselves to the cases where our theories are unitarily equivalent, what can we say about the relation between their vacua $|0\rangle$ and $|\bar{0}\rangle$ other than they are nonorthogonal? First, by inspecting the operator expansions (3.42), one immediately sees that the vacuum states $|0\rangle$ and $|\bar{0}\rangle$ will not in general coincide. For example, we see that:

$$a_i \, |\bar{0}\rangle = \sum_j \beta^*_{ji} \, |\bar{1}_j\rangle \neq 0, \tag{3.52}$$

which yields a nonzero expectation value for particle numbers in a 'mismatched' vacuum, such as:

$$\langle \bar{0} | \, N_i \, |\bar{0}\rangle = \sum_j |\beta_{ji}|^2. \tag{3.53}$$

Such nonzero expectation values come from the fact that generally the annihilation operators from one family mixes creation and annihilation ones from the other, which,

[P] Although we will explicitly calculate them in a special case in section 3.4.2.

in its turn, can be traced back to the β coefficients, which mix the 'not-conjugated' modes (associated with annihilation operators on quantization) with the 'conjugated' modes (associated with creation operators on quantization). In Minkowski spacetime, time-translation symmetry gave a distinct meaning to both subspaces of modes: they were associated with positive and negative frequencies, respectively (see (3.34a)).

A more general class of spacetimes where this priviledged distinction between positive and negative frequency modes arises are stationary spacetimes. These spacetimes will possess (at least one) timelike Killing field ξ^a (see Appendix B) such that, analogously to (3.34a), we can define positive frequency modes u_j as:

$$i\pounds_{\xi} u_j(x) = \omega_j u_j(x)\,, \quad \omega_j > 0 \tag{3.54}$$

(where $\pounds_{\xi}$ denotes the Lie derivative with respect to ξ^a).

However, for general curved spacetimes with no such symmetries, there will be no physically priviledged modes in terms of which to define positive-frequency solutions. In the next section, we shall show that this reflects the fact that the concept of particle as occupation numbers of some field modes has generally no direct physical interpretation in terms of what observers would measure with particle detectors.

3.2.1 Relating mode and Canonical Commutation Relations

Before we proceed to the next section, we shall explicitly show how one may reobtain the canonical commutation relations from (3.30). As in the case of flat space (where we conversely derived (2.62) from (2.54)), the essential factor to this equivalence is that the (classical) maximal information of the field can be extracted both from ϕ and π in a Cauchy surface, or by the (space)time-independent amplitudes α_i and α_i^* for a complete set of modes; correspondingly, in a quantum description one may write the infinite collections of operators $\{\phi(x), \pi(x)\}, x \in \Sigma$ and $\{a_i, a_i^\dagger\}, i \in I$ in terms of one another.

The demonstration will be more convenient if we pick a time coordinate t (in order to define the momentum π) and a foliation Σ_t such that the timelike vector field t^μ whose integral lines generate the evolution in t coincides with n^μ, the unit vectors orthogonal to Σ_t at each event. In this case, we write the metric components in the form:

$$g_{\mu\nu} = n_\mu n_\nu - h_{\mu\nu}, \tag{3.55}$$

with $n_\mu n^\mu = 1$, and being $h_{\mu\nu}$ tangent to each Σ_t (its action restricted to these surfaces defines a positive-definite metric on them), such that $h_{\mu\nu}n^\nu = 0$.

Then, we have the velocity $\dot{\phi} \equiv n^\mu \nabla_\mu \phi \equiv n^\mu \partial_\mu \phi = \partial_0 \phi$. From the Lagrangian (3.23), we obtain the canonically conjugated momentum:

$$\pi \equiv \frac{\partial \mathscr{L}}{\partial \dot{\phi}} = g^{\mu 0} \partial_\mu \phi = n^\mu \partial_\mu \phi. \tag{3.56}$$

Then using the field mode expansions (3.31), and the commutation relations (3.30), we immediately obtain:

$$[\phi(x), \phi(x')] = \sum_i u_i(x) u_i^*(x') - u_i^*(x) u_i(x'), \tag{3.57a}$$

$$[\pi(x), \pi(x')] = \sum_i n^\mu n^\nu \Big(\partial_\mu u_i(x) \partial'_\nu u_i^*(x') - \partial_\mu u_i^*(x) \partial'_\nu u_i(x') \Big), \tag{3.57b}$$

$$[\phi(x), \pi(x')] = \sum_i n^\mu \Big(u_i(x) \partial'_\mu u_i^*(x') - u_i^*(x) \partial'_\mu u_i(x') \Big) \tag{3.57c}$$

(where $\partial'_\mu = \frac{\partial}{\partial x'^\mu}$).

Then, if we wish to analyze these relations for equal times, we must only restrict x and x' to belong to the same Cauchy surface Σ_t, for which we denote $x = (\mathbf{x}, t)$ and $x' = (\mathbf{x}', t)$. Now, by virtue of the completeness of $\{u_i, u_i^*\}$, we should be able to expand any arbitrary solutions $v(x)$ to the field equations (3.25) in terms of it:

$$\begin{aligned} v(x) &= \sum_i (u_i, v) u_i(x) - (u_i^*, v) u_i^*(x) \\ &= \sum_i \int_{\Sigma_t} d\mu_h(\mathbf{x}')\, n^\mu \Big(u_i^*(x') \partial'_\mu v(x') - \partial'_\mu u_i^*(x') v(x') \Big) u_i(x) \\ &\qquad\qquad - n^\mu \Big(u_i(x') \partial'_\mu v(x') - \partial'_\mu u_i(x') v(x') \Big) u_i^*(x) \\ &= \int_{\Sigma_t} d\mu_h(\mathbf{x}') \left[n^\mu \textstyle\sum_i \Big(u_i(x) \partial'_\mu u_i^*(x') - u_i^*(x) \partial'_\mu u_i(x') \Big) \right] v(x') \\ &\qquad\qquad - \left[\textstyle\sum_i \Big(u_i(x) u_i^*(x') - u_i^*(x) u_i(x') \Big) \right] n^\mu \partial'_\mu v(x'), \end{aligned} \tag{3.58}$$

whence we conclude that:

$$\sum_i n^\mu \Big(u_i(x) \partial'_\mu u_i^*(x') - u_i^*(x) \partial'_\mu u_i(x') \Big) = \delta(\mathbf{x}, \mathbf{x}'), \tag{3.59}$$

$$\sum_i \Big(u_i(x) u_i^*(x') - u_i^*(x) u_i(x') \Big) = 0. \tag{3.60}$$

Additionally, taking the time derivative of (3.58), we have:

$$n^{\mu}\partial_{\mu}v(x)=\sum_{i}n^{\mu}\Big((u_i,v)\partial_{\mu}u_i(x)-(u_i^*,v)\partial_{\mu}u_i^*(x)\Big), \tag{3.61}$$

as the inner product is time(Cauchy surface)-invariant. From (3.61), one can analogously derive that:

$$\sum_{i}n^{\mu}n^{\nu}\left(\partial_{\mu}u_i(x)\partial'_{\nu}u_i^*(x')-\partial_{\mu}u_i^*(x)\partial'_{\nu}u_i(x')\right)=0. \tag{3.62}$$

Thus, applying (3.59), (3.60) and (3.62) to (3.57), we immediately recover the canonical commutation relations:

$$[\phi(\mathbf{x},t),\phi(\mathbf{x}',t)]=0=[\pi(\mathbf{x},t),\pi(\mathbf{x}',t)], \tag{3.63a}$$

$$[\phi(\mathbf{x},t),\pi(\mathbf{x}',t)]=i\delta(\mathbf{x},\mathbf{x}'). \tag{3.63b}$$

3.3 Particle Detectors: an empirical notion of particles

From the fact that there are different sets of modes associated with different vacuum states, the question arises of which of these should yield the "most physical vacuum"; that is, loosely speaking, the "most empty" vacuum, or the vacuum that better corresponds to the "experience of no particles". As stated above, this question is notably ill posed, since any empirical notion of "emptiness", or the "experience of no particles", cannot depend on the state of the field alone; at the very least it also requires an observer interacting with it.

And indeed (as we shall see ahead), for a fixed field state, the number of particles measured by an observer will be highly nonunique; among other factors, it will depend on the observer's state of motion. This is true even in Minkowiski spacetime; what is special about the latter is not the existence of a unique vacuum state, but rather that its high degree of symmetry assures that there is a common vacuum state for *all inertial observers.* In general globally hyperbolic spacetimes, no such state will exist (even if we confine ourselves to inertial observers), and there will be an inherent ambiguity in the number of particles measured by different observers, or even by the same observer at different times (while the field remaining in a fixed state).

A great deal of this ambiguity in the concept of particles springs from the fact that they are defined as excitations (occupation numbers) of field modes, which are defined *globally, in the entire spacetime* (indeed, we have for example that a particle with momentum $\mathbf{k}$ will be completely spatially delocalized). This global nature makes it impossible to generally draw simple relations between the expected values $\langle N_i\rangle$ and the (statistical) results of mesurements carried by spatially localized observers (and much less to write simple transformation laws between the results of measurements of 2 distinct observers). In contrast, local observables such as $\langle T_{\mu\nu}(x)\rangle$ allow a more direct interpretation in terms of measurements carried by localized observers. Furthermore, they are subject to simple transformation laws relating what is measured by two different observers; in the specific case of a tensorial quantity, such as the stress tensor $\langle T_{\mu\nu}(x)\rangle$, this relation should be a simple coordinate transformation relating two reference frames. Particularly, if $\langle T_{\mu\nu}(x)\rangle = 0$ for one observer, the same should be true for *all* observers.

Still, in a few highly symmetric spacetimes, a privileged notion of particles may arise, which will be associated with special modes (following spacetimes symmetries and being related to special families of observers). In such cases, simple relations will emerge between the expected values $\langle N_i\rangle$ and the particles measured by these special observers, recovering, for example, the well-known particle notion in Minkowski spacetime. One particular case of interest is that of spacetimes which are assimptotically Minkowskian in the remote past and remote future: in this case, both regions will have special vacuum states, which we respectively denote as $|0_p\rangle$ and $|0_f\rangle$.

We remind that, as we are working in the Heisenberg picture, the vector states remain unchanged under time evolution, and the same is true for any number operators $N_{\mathbf{k}}$, as they are defined in respect to global modes, defined through all of spacetime. However, the *physical* notion of particles, as measured by particle observables (and particularly *which* set of number operators may be of direct physical significance) will generally change with time.

We shall illustrate all of the above considerations by exploring an idealized model of a particle detector[Q]. This model consists of a point-like physical system, whose internal degrees of freedom correspond to a discrete set of energies $\{E\}$, whose internal dynamics are given by the Hamiltonian H_0: $H_0\,|E\rangle = E\,|E\rangle$. For simplicity, we take its energy levels to be nondegenerate (*i.e.* we take H_0 to be a C.S.C.O. for the system).

Now, this probe system (the 'detector') shall be weakly coupled to our scalar field by a local monopole interaction, given by Lagrangian:

$$\mathscr{L}_I = c\,m(\tau)\phi(x^\mu(\tau)), \tag{3.64}$$

where τ denotes the proper time of the detector, $x^\mu(\tau)$ its (classical) trajectory in spacetime, $m(\tau)$ its monopole moment, and c a "small" coupling constant.

Within this framework, we are interested in deriving the probabilities that the interaction with the field will promote a "detection", that is, an excitation of our probe

[Q] Here, we follow (and extend a little) the exposition in (1). See section 3.3 on it, and references therein.

system from its ground state $|E_0\rangle$ to an excited state $|E\rangle$, $E > E_0$. We are demanding the coupling to be weak, so that the interactions between the field and detector may be treated perturbatively. Formally, we shall derive the probabilities of excitation of the detector like a scattering problem[R], through the S matrix formalism[S]. We thus switch from the Heisenberg to the Dirac picture, where the field and detector observables evolve through their free Hamiltonians, whereas states evolve through the interaction Hamiltonian $\mathcal{H}_i = -\mathscr{L}_I$. To first perturbative order, this entails the transition amplitudes $\mathcal{A}$ between two states $|E, \Psi\rangle$ and $|E', \Psi'\rangle$:

$$\mathcal{A}\Big(|E,\Psi\rangle \to |E',\Psi'\rangle\Big)= ic\,\langle E',\Psi'|\int_{-\infty}^{+\infty} m(\tau)\phi(x^\mu(\tau))d\tau\,|E,\Psi\rangle\,. \tag{3.65}$$

Particularly, we are interested in the possibility of making a transition to make a detection in the vacuum state. That is, of starting with our field in a vacuum $|0\rangle$ and our detector at ground state $|E_0\rangle$, and ending up with an excited detector $|E\rangle$ and some final state for the field $|\Psi\rangle$ (the precise state $|\Psi\rangle$ of the field *after* the measurement is of little importance to us; the relevant question here is if we can make a detection). Let us then analyze the probabilities of detection in Minkowski space, in the usual Minkowski vacuum $|0_M\rangle$:

$$\mathcal{A}\Big(|E_0,0_M\rangle \to |E,\Psi\rangle\Big)= ic\,\langle E,\Psi|\int_{-\infty}^{+\infty} m(\tau)\phi(x^\mu(\tau))d\tau\,|E_0,0_M\rangle\,. \tag{3.66}$$

Since we are working in the Dirac Picture, $m(\tau)$ simply evolves through the free Hamiltonian H_0:

$$m(\tau) = e^{iH_0\tau}m(0)e^{-iH_0\tau}. \tag{3.67}$$

Substituting in (3.66), we obtain:

$$\mathcal{A}(|E_0,0_M\rangle \to |E,\Psi\rangle) = ic\,\langle E|\,m(0)\,|E_0\rangle \int_{-\infty}^{+\infty} e^{i(E-E_0)\tau}\,\langle\Psi|\,\phi(x^\mu(\tau))\,|0_M\rangle\,d\tau. \tag{3.68}$$

Since ϕ is linear in creation and annihilation operators, the only transitions that may occur (*in first perturbative order*) are those to one-particle states: $|\Psi\rangle = |1_{\mathbf{k}}\rangle$. If we consider the continuum normalization (2.36), we get the amplitudes:

[R] Note that this entails the assumption that interactions are transient. However, as one can easily see from (3.64), the interactions are generally persistent. This will lead to a few incongruences below, which will be addressed in due time.

[S] The reader unfamiliar with this formalism is referred to chapter 6 of (6) for further details.

$$\begin{aligned}\langle 1_{\mathbf{k}}|\,\phi(x)\,|0\rangle &= \int d^3\mathbf{k}'(16\pi^3\omega_{k'})^{-1/2}\,\langle 1_{\mathbf{k}}|\,a^\dagger_{\mathbf{k}'}\,|0\rangle\, e^{i\omega' t - i\mathbf{k}'\,\cdot\,\mathbf{x}} \\ &= (16\pi^3\omega_k)^{-1/2} e^{i\omega t - i\mathbf{k}\,\cdot\,\mathbf{x}}. \end{aligned} \tag{3.69}$$

Inserting the result in (3.68), we see that we must indeed specify a spacetime trajectory $x^\mu(\tau)$ to the detector to compute well-defined transition amplitudes. Let us first consider an inertial world-line:

$$\mathbf{x} = \mathbf{x_0} + \mathbf{v}t = \mathbf{x_0} + \mathbf{v}\gamma_v\tau, \tag{3.70}$$

where γ_v is the Lorentz factor $\gamma_v = (1 - v^2)^{-\frac{1}{2}}$. In this case, we have:

$$\begin{aligned}\mathcal{A}(|E_0, 0_M\rangle \to |E, 1_{\mathbf{k}}\rangle) &= \frac{ic\,\langle E|\,m(0)\,|E_0\rangle}{16\pi^3\omega} e^{-i\mathbf{k}\,\cdot\,\mathbf{x_0}} \int_{-\infty}^{+\infty} e^{i(E-E_0)\tau} e^{i(\omega - \mathbf{k}\,\cdot\,\mathbf{v})\gamma_v\tau} d\tau \\ &= \frac{ic\,\langle E|\,m(0)\,|E_0\rangle}{4\pi\omega} e^{-i\mathbf{k}\,\cdot\,\mathbf{x_0}} \delta(E - E_0 + [\omega - \mathbf{k}\cdot\mathbf{v}]\gamma_v). \end{aligned} \tag{3.71}$$

But since $E > E_0$ and $\omega > |\mathbf{k}\cdot\mathbf{v}|$ (as $v < 1$ for any timelike trajectory and $\omega = \sqrt{k^2 + m^2} \geq k$) there are no roots in the arguments of the δ distribution in (3.71), and the transition amplitude is always zero, as dictated by energy conservation – a direct consequence of time translation symmetry (as energy is the global Noether charge associated to this symmetry).

For more complicated trajectories, however, the transition amplitudes (3.68) do not generally yield δ's, and nonzero transition probabilities may emerge from the vacuum! (As we shall demonstrate briefly.) In such cases, we will be interested in summing the transition probabilities over all possible final states $|\Psi\rangle$ and $|E\rangle$ ($\neq |E_0\rangle$) to obtain the total probability that *any* transition (detection) may occur:

$$\begin{aligned}\sum_{E,\Psi}\Big|\mathcal{A}(|E_0, 0_M\rangle \to |E, \Psi\rangle)\Big|^2 = c^2 \sum_E \Big\{ &|\langle E|m(0)|E_0\rangle|^2 \times \\ &\iint d\tau\, d\tau' e^{i(E-E_0)(\tau-\tau')}\,\langle 0_M|\phi(\tau')[\textstyle\sum_\Psi |\Psi\rangle\langle\Psi|]\phi(\tau)|0_M\rangle \Big\}. \end{aligned} \tag{3.72}$$

Using the completeness relation $\sum_\Psi |\Psi\rangle\langle\Psi| = \mathbb{1}$, and recognizing the vacuum two-point correlation as the Wightman function (2.143), we have:

$$\begin{aligned} P &= c^2 \sum_E |\langle E|m(0)|E_0\rangle|^2 \iint d\tau d\tau' e^{-i(E-E_0)(\tau-\tau')} G^+(x(\tau), x(\tau')) \\ &= c^2 \sum_E |\langle E|m(0)|E_0\rangle|^2 \mathscr{F}(E - E_0), \end{aligned} \tag{3.73}$$

where we defined the response function of the detector $\mathscr{F}(E)$:

$$\mathscr{F}(E) \equiv \iint d\tau\, d\tau' e^{-iE(\tau-\tau')} G^{+}(\tau, \tau'). \tag{3.74}$$

(Here, we simplified the notation of G^{+}, leaving implicit the dependence on the detector trajectory $x(\tau)$.)

Taking a closer look at expression (3.73), we see that the details regarding the inner structure of the detector enter only in the prefactor $c^2 |\langle E|m|E_0\rangle|^2$, whereas the response function carries the dependence on the field variables (of course, it will also depend on the detector energy differences, just like the response of an atom interacting with radiation will depend on its spectrum). If we are not particularly interested in this inner structure, but rather in the field-related response, we may just focus on the latter.

Then, to evaluate (3.74) more closely, it is convenient to perform change of variables in this double integral, analyzing it in terms of the time average $\bar{\tau} = \frac{1}{2}(\tau + \tau')$ and time difference $\Delta\tau = \tau - \tau'$. Since the transformation $(\tau, \tau') \to (\bar{\tau}, \Delta\tau)$ has unit Jacobian, we have:

$$\mathscr{F}(E) = \iint d\bar{\tau}\, d(\Delta\tau) e^{-iE\Delta\tau} \tilde{G}^{+}(\bar{\tau}, \Delta\tau), \tag{3.75}$$

where $\tilde{G}^{+}(\bar{\tau}, \Delta\tau) \equiv G(\tau, \tau')$.

Particularly, if we analyze a *stationary* trajectory, that is, one for which the correlations $G^{+}(\tau, \tau')$ only depend on the proper-time *differences*, $G^{+}(\tau, \tau') = G^{+}(\Delta\tau)$ (here, we drop the tilde in our notation since there is no risk of ambiguity), we obtain trivially separable integrals:

$$\mathscr{F}(E) = \left(\int_{-\infty}^{\infty} d\bar{\tau}\right)\left(\int_{-\infty}^{\infty} d(\Delta\tau) e^{-iE\Delta\tau} G^{+}(\Delta\tau)\right), \tag{3.76}$$

which can be immediately interpreted as a (constant) *transition rate* multiplied by the (infinite) time interval of the interactions $T \equiv \int d\bar{\tau}$. It is clear that, whenever we have non null transition rates, such a case will yield divergent transition probabilities. This evidently points to a break in our perturbative approximation for indefinitely long time scales with persistent interactions, as anticipated earlier; this occurs because a (first-order) perturbative approach fails to account for the possibility that the system already transitioned in a past instant, acumulating an (unboundedly) increasing transition probability.

Nonetheless, this approach should render a good approximation if we restrict our analysis to sufficiently short time intervals T, for which $\mathscr{F}(E) \ll 1$, that is:

$$T \ll \left(\int_{-T}^{T} d(\Delta\tau) e^{-iE\Delta\tau} G^{+}(\Delta\tau)\right)^{-1}, \tag{3.77}$$

but sufficiently long so that there will not be a great difference in setting the integration limits at this finite T, rather than at infinity – note that the faster the vacuum correlations $G^{+}(\Delta\tau)$ decay, and the higher the energy jump E in the detector is, the smaller this lower bound will be (and the greater the upper bound will be). (Physically, one can think of this restriction as considering a detector that does not eternally interact with the "background" field, but rather that is set to interact with it for a finite time interval T.[T])

As long as we remain in these consistency intervals, it is actually quite more convenient to work directly with *transition rates*. Thus, we define the *response function per unit time*:

$$\mathscr{F}'(E) = \frac{\mathscr{F}(E)}{T} = \int_{-\infty}^{\infty} d(\Delta\tau) e^{-iE\Delta\tau} G^{+}(\Delta\tau). \tag{3.78}$$

Let us then attempt to evaluate this function explicitly. Even with all the simplifications so far, Green function G^{+} is still a little convoluted to resolve analytically in the massive case, $m > 0$. Thus, we restrict our attention to the simpler case of a massless field, $m{=}0$, and analyze it in further detail. In this case, for an arbitrary pair of events (x, x'), G^{+} reads:

$$\begin{aligned} G^{+}(x,x') &= \frac{-i}{(2\pi)^4}\int d^4k \frac{e^{-ik(x-x')}}{(k^0)^2 - \mathbf{k}^2} \\ &= \frac{1}{(2\pi)^3}\int \frac{d^3\mathbf{k}}{2|\mathbf{k}|} e^{-i|\mathbf{k}|\Delta t + i\mathbf{k}\cdot\Delta\mathbf{x}} \\ &= \frac{1}{(2\pi)^3}\int_0^{\infty} \frac{d|\mathbf{k}|}{2|\mathbf{k}|}|\mathbf{k}|^2 e^{-i|\mathbf{k}|\Delta t}\int_{-1}^{1} d(\cos\theta) e^{i|\mathbf{k}||\Delta\mathbf{x}|\cos\theta}\left(\int_0^{2\pi} d\phi\right) \\ &= \frac{1}{4\pi^2}\frac{1}{2i|\Delta\mathbf{x}|}\int_0^{\infty} d|\mathbf{k}|(e^{-i|\mathbf{k}|(\Delta t - \Delta x)} - e^{-i|\mathbf{k}|(\Delta t + \Delta x)}). \end{aligned} \tag{3.79}$$

This integral obviously does not converge in the usual functional sense. As we have seen in Section 2.5 and Appendix A, we must generally interpret two-point functions in integrals such as (3.73) in the distributional sense.

However, a convenient trick to work directly with G^{+} (*i.e.* to get a closed expression for G^{+}, carrying the k-integral (3.79) *before* the $\Delta\tau$ integral in (3.78)) is to introduce the regularizer $e^{-\epsilon|\mathbf{k}|}$ ($\epsilon > 0$), making (3.79) absolutely convergent. In the end of *all*

[T] In this case, one requires the "offswitch" (decoupling) of the detector to occur adiabatically (in a sufficiently smooth and slow manner so that no particles are created by the process).

integrations, we may relax the regularization and take the limit $\epsilon \to 0^+$. Denoting this regularized function by G^+_ϵ, we have:

$$\begin{aligned} G^+_\epsilon(x, x') &= \frac{1}{4\pi^2}\frac{1}{2|\Delta\mathbf{x}|}\left(\frac{1}{\Delta t - i\epsilon - |\Delta\mathbf{x}|} - \frac{1}{\Delta t - i\epsilon + |\Delta\mathbf{x}|}\right) \\ &= \frac{1}{4\pi^2}\frac{1}{(\Delta t - i\epsilon)^2 - |\Delta\mathbf{x}|^2}. \end{aligned} \tag{3.80}$$

In the case of an inertial detector (3.70), we have:

$$\frac{1}{(\Delta t - i\epsilon)^2 - |\Delta\mathbf{x}|^2} = \frac{1}{(\gamma_v\Delta\tau - i\epsilon)^2 - (\gamma_v v\Delta\tau)^2} = \frac{1}{\Delta\tau^2 - 2i\Delta\tau\gamma_v\epsilon + \mathcal{O}(\epsilon^2)}.$$

We then absorb the positive factor γ into ϵ and ignore any higher order ($\mathcal{O}(\epsilon^2)$) corrections to write:

$$G^+_\epsilon(x, x') = \frac{1}{4\pi^2(\Delta\tau - i\epsilon)^2}. \tag{3.81}$$

Substituting this in the integral (3.78), we can easily compute it as a contour integral, invoking Cauchy theorem. For $E > 0$, we should close the integration contour at the lower half of the complex plane. Then, since the only pole of the integrand lies in the upper plane, at $\Delta\tau = +i\epsilon$, we have that the response rate of the detector is null. Transporting this result to (3.73), with $E - E_0 > 0$, obtain a null detection probability, in perfect accordance with our previous result (3.71).

However, even in the simple situation of stationary response in Minkowski spacetime, we can still find nontrivial examples of particle detection. A case of particular interest is a uniformly accelerated detector, with constant proper acceleration $a = \alpha^{-1}$. Such a detector describes a hyperbolic trajectory in spacetime, which may be conveniently described by *inertial coordinates* in the xt-plane as:

$$\begin{cases} x(\tau) = \alpha\cosh(\tau/\alpha) & (3.82a) \\ t(\tau) = \alpha\sinh(\tau/\alpha) & (3.82b) \end{cases}.$$

By substituting (3.82) in (3.79), one finds (with some algebraic effort) that:

$$G^+_\epsilon(\Delta\tau) = \left[16\pi^2\alpha^2\sinh^2\!\left(\frac{\Delta\tau - 2i\epsilon}{2\alpha}\right)\right]^{-1}, \tag{3.83}$$

where we have once again absorbed a finite positive factor, $f(\tau, \tau')$, into ϵ, given by:

$$f(\tau,\tau') \equiv \frac{\sinh(\tau/\alpha)-\sinh(\tau'/\alpha)}{\sinh((\tau-\tau')/\alpha)} > 0, \quad \forall \tau,\tau'. \tag{3.84}$$

Then, substituting (3.83) in (3.78), we can once again compute the integral through Cauchy Theorem, in a conveniently chosen contour (see Figure 17). Note that G^+_ϵ is periodic along the imaginary axis, and its poles lie regularly at $z = 2i(\epsilon + n\pi\alpha)$. Then, by closing the contour rectangularly after one period, as illustrated in the figure, and denoting the *regularized* integral in the real axis as I_ϵ, we get:

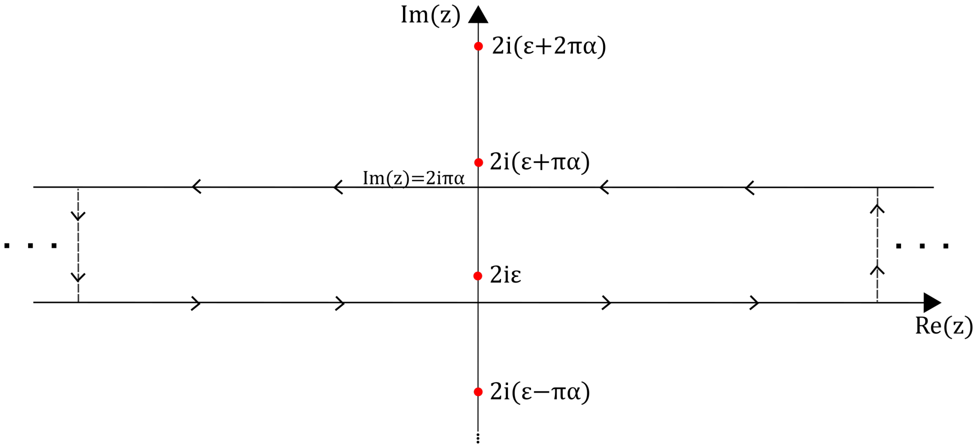

Figure 17 – Illustration of the contour integral used to calculate the reponse rate for the accelerated detector. The poles of the integrand lie along the imaginary axis, and are represented by red dots. The chosen contour is shown in thick black lines (the dashed lines that close the rectangle should be considered at infinity, where they will give no contribution to the integral), and it only encompasses the pole at $z = 2i\epsilon$.
Source: By the author.

$$(1 - e^{2\pi\alpha(E-E_0)})I_\epsilon = 2\pi i \operatorname*{Res}_{z=2i\epsilon}\Big(e^{-iEz}G^+(z)\Big). \tag{3.85}$$

This residue may be calculated for this order 2 pole by:

$$\operatorname*{Res}_{z=2i\epsilon}(g) = -\frac{1}{4\pi^2}\lim_{z\to i\epsilon}\frac{d}{dz}\left[\frac{(\frac{z-2i\epsilon}{2\alpha})^2}{\sinh^2(\frac{z-2i\epsilon}{2\alpha})}e^{-i(E-E_0)z}\right] = \frac{i}{4\pi^2}(E-E_0)e^{2\epsilon(E-E_0)}, \tag{3.86}$$

so that we finally obtain the transition probability rate per unit time, $P' = P/T$, in (3.73):

$$P' = \lim_{\epsilon\to 0^+} c^2 \sum_E |\langle E|m(0)|E_0\rangle|^2 I_\epsilon = \frac{c^2}{2\pi}\sum_E \frac{|\langle E|m(0)|E_0\rangle|^2}{e^{2\pi(E-E_0)\alpha}-1}. \tag{3.87}$$

Upon immediate inspection of this transition rate, we identify a Planck factor $(e^{2\pi(E-E_0)\alpha}-1)^{-1}$, showing that this detection rate corresponds to a thermal distribution of particles, with an effective temperature proportional to the proper acceleration of the detector: $T = (2\pi\alpha)^{-1} = a/2\pi$.

Note, moreover, that a transition will generally excite *both the detector and the field*, being the final state $|\Psi\rangle$ of the latter generally a 1-particle state. At first sight, this seems very strange on energy grounds, since both the field and the detector will raise their energy (and, in this example, we remain in Minkowski spacetime, whose time translation symmetry should enforce energy conservation). How, then, can we reconcile these nonvanishing transition rates with energy conservation? As it turns out, we ought to attribute the injection of energy in our system to the agent that is imprinting acceleration in our detector: as the latter is coupled to ϕ, it should indeed cause the emission of particles whenever it is under accelerated motion, imposing upon it a 'breaking' force in the opposite direction of the acceleration (this is analogous to electromagnetic *Brehmstrahlung*, where an accelerated charge emits radiation). Thus, the external agent that maintains an acceleration on the coupled ("charged") detector must do work on it against this breaking force, providing energy for both the excitation of the detector and the emission of particles.

Now that we have seen a nontrivial application of particle detection in the vacuum, let us show next that our detector model indeed reproduces the expected results of particle detection for inertial observers in Minkowski space. These will *exceptionally* bear a simple relation with the expected values $\langle N_{\mathbf{k}}\rangle$ for occupation numbers in plane-wave modes. We begin by analyzing the response rate (3.78) for general many-particle states $|\Psi\rangle = |n_{\mathbf{k_1}}, n_{\mathbf{k_2}}, ...\rangle$. In this case, we must substitute the Wightman function G^+ by:

$$\begin{aligned} G^+(x,x') \equiv \langle 0_M|\phi(x)\phi(x')|0_M\rangle \longrightarrow &\langle\Psi|\phi(x)\phi(x')|\Psi\rangle \\ &= G^+(x,x') + \sum_{\mathbf{k}} n_{\mathbf{k}}\Big(u_{\mathbf{k}}(x)u^*_{\mathbf{k}}(x') + u^*_{\mathbf{k}}(x)u_{\mathbf{k}}(x')\Big), \end{aligned} \tag{3.88}$$

or, taking the continuum limit:

$$\langle\Psi|\phi(x)\phi(x')|\Psi\rangle = G^+(x,x') + \int d^3\mathbf{k}\, n(\mathbf{k})u_{\mathbf{k}}(x)u^*_{\mathbf{k}}(x') + \int d^3\mathbf{k}\, n(\mathbf{k})u^*_{\mathbf{k}}(x)u_{\mathbf{k}}(x'). \tag{3.89}$$

This expression gives us three contributions for the response rate. However, we already know that the first term corresponds to the vacuum contribution, which yields a null response for an inertial observer. Let us then look at the contributions from the second and third terms of (3.89) in (3.78):

$$\begin{cases} \dfrac{1}{(2\pi)^3}\displaystyle\int \frac{d^3\mathbf{k}}{2\omega} n(\mathbf{k}) \int\limits_{-\infty}^{\infty} d(\Delta\tau) e^{-i[E+\gamma_v(\omega-\mathbf{k}\,\cdot\,\mathbf{v})]\Delta\tau} = \dfrac{1}{(2\pi)^3}\displaystyle\int \frac{d^3\mathbf{k}}{2\omega} n(\mathbf{k}) \delta\Big(\cancelto{0}{E+\gamma_v(\omega-\mathbf{k}\cdot\mathbf{v})}\Big) \\ \dfrac{1}{(2\pi)^3}\displaystyle\int \frac{d^3\mathbf{k}}{2\omega} n(\mathbf{k}) \int\limits_{-\infty}^{\infty} d(\Delta\tau) e^{-i[E-\gamma_v(\omega-\mathbf{k}\,\cdot\,\mathbf{v})]\Delta\tau} = \dfrac{1}{(2\pi)^3}\displaystyle\int \frac{d^3\mathbf{k}}{2\omega} n(\mathbf{k}) \delta\Big(E-\gamma_v(\omega-\mathbf{k}\cdot\mathbf{v})\Big) \end{cases} .$$

(3.90a)

(3.90b)

Note that the δ in (3.90a) has no roots for $E > 0$. The one in (3.90b), however, has roots in the domain of integration, and will yield a nonnull contribution to the response rate. We can already see from this expression that, generally, this contribution for an excitation ΔE in our detector will come precisely from particles that have energy equal to ΔE *as seen in the detector reference frame*. Let us compute this response rate explicitly for an isotropic particle distribution (*i.e.* $n(\mathbf{k}) = n(k)$) and a detector at rest in the isotropic frame ($\mathbf{v} = 0$):

$$\begin{aligned} \mathscr{F}'(E) &= \frac{1}{(2\pi)^3}\Big(\int d\Omega\Big)\int\limits_0^{\infty} \frac{dk}{2\omega} k^2 n(k)\delta(E-\omega) \\ &= \frac{1}{4\pi^2}\int\limits_m^{\infty} d\omega \sqrt{\omega^2-m^2}\,\bar{n}(\omega)\delta(E-\omega) \\ &= \frac{1}{4\pi^2}\sqrt{E^2-m^2}\,\bar{n}(E)\Theta(E-m), \end{aligned} \tag{3.91}$$

where we have defined the energy particle distribution $\bar{n}(\omega) \equiv n(\sqrt{\omega^2-m^2}) = n(k)$.

The results in (3.91) are quite straightforward to interpret. The detection rates at an energy E are proportional to the particle density at that energy (and a factor accounting for the spectral surface area k^2 divided by the k-dependent normalization), and the Heaviside Θ (re)assures that one can make detections only above the minimum threshold energy, given by $\omega = m$.

For an anysotropic observer ($\mathbf{v} \neq 0$), the response function does not look as simple due to a Doppler spreading, but the results are also easy to account for:

$$\mathscr{F}(E) = \frac{1}{(2\pi)^3}\Big(\int\limits_0^{2\pi} d\varphi\Big)\int\limits_0^{\infty} \frac{dk}{2\omega} k^2 n(k) \int\limits_{-1}^{+1} d(\cos(\theta))\delta\Big(\cancelto{\text{Has roots if } \omega - kv < \frac{E}{\gamma_v} < \omega + kv}{E-\gamma_v(\omega-kv\cos(\theta))}\Big)$$

$$
\begin{aligned}
&= \frac{1}{4\pi^2}\frac{1}{\gamma_v v}\int\limits_0^\infty \frac{dk}{2\omega} k n(k)\Big[\Theta(\omega - E_-) - \Theta(\omega - E_+)\Big] \\
&= \frac{1}{4\pi^2}\left(\frac{1-v^2}{v^2}\right)^{1/2}\int_{E_-}^{E_+} d\omega\ \bar{n}(\omega),
\end{aligned}
\tag{3.92}
$$

where:

$$
E_\pm = \frac{E \pm \sqrt{(E^2 - m^2)v^2}}{\sqrt{1-v^2}}. \tag{3.93}
$$

Again, we find that nonnull transition rates are only possible for $E > m$. This formula is particularly simple in the massless case, where all the energy of the particles comes from a kinetic term. In this case, the Doppler-shifted energies are just:

$$
E_\pm = E\frac{1 \pm v}{(1-v^2)^{\frac{1}{2}}} = E\left[\frac{1 \pm v}{1 \mp v}\right]^{\frac{1}{2}}. \tag{3.94}
$$

To wrap-up the discussion in this section: we have taken a closer look at the ambiguities in the concept of particles and, with the aid of simple models of particle detectors, we have seen how such ambiguities reflect in highly nontrivial observation relations for particles, even in (what should be arguably the most trivial and devoid of all states:) the Minkowski vacuum. Already through these examples, we may glimpse that *there is in general no simple relation between the expected value $n_i \equiv \langle\Psi|N_i|\Psi\rangle$ and the number of particles measured by an actual detector, even for an inertial (free-falling) detector*[U].

However, we have shown that *in the very particular case* of free-falling detectors in Minkowski spacetime, this simple relation does exist and the operational definition of particles constructed from detectors coincides with that given by the populations of normal (plane-wave) modes. In the next section, we will be interested in a less trivial context for which the normal-mode definition of particles is still useful – namely, spacetimes with transient dynamics –, and will allow us to define the phenomenon of particle creation in dynamical spacetimes.

[U] In some instances, as in the case of the accelerated observer in Minkowski, one can actually build appropriate accelerated modes, in terms of which the Minkowski vacuum is a thermal distribution of accelerated particles. This construction, however, relies on an entire family of accelerated observers covering (a wedge of) spacetime, whereas we are interested in the response of just individual localized detectors; this brings us back to the matter that *modes are defined globally*, and, generally, they will not bear a simple relation to locally measured quantities.

3.4 Particle Creation in Asymptotically Flat Spacetimes

In the last section, we have seen that only in very special cases we will find a simple correspondence between the idealized concept of particles as occupation numbers of normal modes, and the empirical one of particles as what particle detectors measure. This correspondence will be possible only when there is a high degree of symmetry in the spacetime under consideration, which picks out both special families of normal modes and special families of observers. Particularly, *time translation symmetry* (and thus stationary spacetimes) plays a proeminent role, since it allows one to define positive-frequency modes (see eq. 3.54) and gives a very simple class of special observers, namely, *stationary observers* (*i.e.* the ones whose worldlines coincide with the orbits of the time translation Killing field ξ^μ).

Now, in order to analyze some effects in dynamical (nonstationary) spacetimes but still keep things simple enough so that the mode particle definition is still useful to draw simple predictions, we turn our attention to the slightly more general case of spacetimes which go through a dynamical period, but that are asymptoptically stationary in the remote past and future; we denote such asymptotic regions Ω_p and Ω_f, respectively. In these spacetimes, both regions will have special sets of normal modes associated with them (which we denote $\{u_j^{(p)}\}$ and $\{u_j^{(f)}\}$, respectively) whose asymptotic behaviour will be of the form[V]:

$$\begin{cases} u_i^{(p)}(x) \simeq \dfrac{e^{-i\omega_i t}}{\sqrt{2\omega_i}}\psi_i^{(p)}(\mathbf{x}), \qquad x \in \Omega_p & (3.95a) \\ u_j^{(f)}(x) \simeq \dfrac{e^{-i\omega_j t}}{\sqrt{2\omega_j}}\psi_j^{(f)}(\mathbf{x}), \qquad x \in \Omega_f & (3.95b) \end{cases},$$

i.e. they approximate positive-frequency modes in their respective asymptotic regions. We stress that both (sets of) modes are defined *in the entire spacetime*, as they are *exact* solutions to the field equations everywhere. However their form outside of their respective asymptotic regions is generally quite complicated and will depend heavily on the spacetime evolution.

As discussed before, we may write field expansions in both mode sets:

$$\phi(x) = \sum_i a_i^{(p)} u_i^{(p)}(x) + a_i^{\dagger(p)} u_i^{*(p)}(x) = \sum_j a_j^{(f)} u_j^{(f)}(x) + a_j^{\dagger(f)} u_j^{*(f)}(x), \tag{3.96}$$

[V] From this point onward, we shall always denote the set of spatial coordinates $\{x^j\}_{j=1,2,3}$ by boldface letters, as we do for ordinary spatial vectors in $\mathbb{R}^3$, even though we are not necessarily considering spatially flat Cauchy surfaces. We do so to compactly distinguish it from 4-dimensional spacetime coordinates/events, which we shall denote just as $x = (t, \mathbf{x})$.

and define number operators for each of them, $N_i^{(p)} \equiv a_i^{\dagger(p)} a_i^{(p)}$ and $N_j^{(f)} \equiv a_j^{\dagger(f)} a_j^{(f)}$, as well as their respective vacuum states $|0_p\rangle$ and $|0_f\rangle$.

It will be of special interest to us when the regions Ω_p and Ω_f are asymptotically *Minkowskian*, in which case the modes $\psi^{(p)}$ and $\psi^{(f)}$ will be just ordinary plane waves, which are particularly simple to operate with.

Since in this simple case one can ascribe a very clear physical meaning to the expected values $\langle\Psi|N_i^{(p)}|\Psi\rangle$ and $\langle\Psi|N_j^{(f)}|\Psi\rangle$ in terms of particles measured by inertial detectors in either the far past or future, one can then refer to the phenomenon of particle creation (or annihilation) between these two regions by means of simple Bogolubov transformations. For example, if we consider our field to be in the vacuum state $|0_p\rangle$, inertial observers in the far past ($x \in \Omega_p$) would indeed measure no particles $\langle 0_p|N_i^{(p)}|0_p\rangle = 0$, $\forall i$. However, *after the time evolution through the dynamical region*[W], inertial observers will generally measure a nontrivial particle content at late times ($x \in \Omega_f$), given by eq (3.52):

$$\langle 0_p|N_j^{(f)}|0_p\rangle = \sum_i |\beta_{ji}|^2, \tag{3.97}$$

being $\beta_{ji} = -(u_j^{(p)}, u_i^{*(f)})$ the usual β Bogolubov coefficients between past and future modes.

Of course, one could in principle also have a symmetrical situation of particle annihilation, starting from the *final* vacuum state $|0_f\rangle$, for which one would measure particles in the past,

$$\langle 0_f|N_j^{(p)}|0_f\rangle = \sum_i |\beta_{ij}|^2, \tag{3.98}$$

but none in the future. However, although this situation is perfectly compatible with an idealized time evolution of a pure state, *it corresponds to a diminishing in enytropy.* As we have commented in section 3.2 (and we shall show more explicitly for FLRW spaces in this section), particles are always created in correlated pairs. This means that a state that evolves from many particles into a vacuum would correspond to an initial state of highly correlated particles, that are perfectly adjusted to be annihilated in pairs (this would be analogous, for example, to postulating an extremely fine-tuned choice of initial conditions for molecules of gas in a box, allowing one to evolve from a state in which the gas is filling the entire box to one in which it spontaneouly concentrates in a fraction of its volume).

So far, the analysis seems quite simple. In practice, however, it is generally quite complicated to actually solve the field equations *exactly* in such generic spacetimes and

[W] Recall that this evolution leaves the state $|0_p\rangle$ unchanged in the Heinsenberg picture.

properly combine a basis of *exact* solutions to obtain the modes whose asymptotic behaviour is that of plane waves in either remote region, as well as to further calculate the Bogolubov coefficients α_{ij} and β_{ij} to *every* pair of modes $u_i^{(p)}$ and $u_j^{(f)}$. Notwithstanding, there is a particular class of spacetimes for which these calculations are greatly simplified: spatially homogeneous and isotropic universes. Throughout this section, we shall explore them as a tractable case of study, and analyze the phenomenon of particle creation in more detail.

3.4.1 Particle creation in FLRW spacetimes

A very distinguished class of spacetimes, which is of special interest in the context of cosmology, are the ones which possess maximally symmetrical space sections, *i.e.* which are spatially homogeneous and isotropic (but which may still have a nontrivial time evolution). For historical reasons, they are also known as *Friedmann-Lemaitre-Robertson-Walker* (FLRW) spaces or universes (for a more complete account of the development and properties of FLRW spaces, as well as their use in cosmology, see section 5.1.1). All spacetimes in this class may be described by a metric of the form:

$$ds^2 = dt^2 - a^2(t)d\Sigma^2, \tag{3.99}$$

where t represents the proper time of observers whose worldlines are orthogonal to the isotropic space sections Σ_t (which foliate the entire spacetime). Such observers, commonly called *comoving observers*, for reasons to be made apparent, comprise a special family in FLRW spaces, as they are the ones who will perceive space (*i.e.* their spatial sections Σ_t) as homogeneous and isotropic. $d\Sigma^2$ represents a static spatial metric (common to all surfaces Σ_t) and $a(t)$ is called the *scale factor*; it dictates how spatial distances expand or shrink with time (*e.g.* $a(t')/a(t)$ gives the ratio of the distances between 2 comoving observers measured along the surfaces $\Sigma_{t'}$ and Σ_t). Particularly, for a FLRW spacetime that is asymptotically static, we must have that:

$$\begin{cases} a(t) \to a_1, \qquad t \to -\infty & \text{(3.100a)} \\ a(t) \to a_2, \qquad t \to +\infty & \text{(3.100b)} \end{cases}$$

(being a_1 and a_2 constants).

What makes these spacetimes special in the context of particle creation is that they bear separable field equations *at all times*, so that one may always find a complete set of field solutions of the form:

$$u_i(x) = \chi_i(t)\psi_i(\mathbf{x}). \tag{3.101}$$

Presently, we shall not go into detail for the dynamical equations (these will be further developed for a conformal time coordinate in section 3.5, and in proper-time in section 4.2). We just note here that, by defining $h_i(t) \equiv a^{-\frac{3}{2}}(t)\chi(t)$ they will take the general form:

$$\begin{cases} \dfrac{d^2}{dt^2}h_i(t) = -\Omega_i^2(t)h_i(t) & (3.102a) \\ H_x^2\psi_i(\mathbf{x}) = \Omega_i^2(t)\psi_i(\mathbf{x}) & (3.102b) \end{cases} ,$$

where $H_x^2 = \left[-a^{-2}(t)\nabla_{\mathbf{x}}^2 + m^2\right]$ (being $\nabla_{\mathbf{x}}^2$ the Laplacian operator corresponding to the metric $d\Sigma^2$), and $\Omega_i(t)$ a *time-dependent frequency.* Since we are particularly interested in asymptotically Minkowskian spaces, we shall restrict ourselves to the case of spatially flat homonegeous surfaces $\Sigma_t = \mathbb{R}^3$. In this case, $\nabla_{\mathbf{x}}^2$ is an ordinary 3D Laplacian and we have simple exponential solutions, labeled by a wave-vector $\mathbf{k}$:

$$\psi_{\mathbf{k}}(\mathbf{x}) = \frac{1}{\sqrt{V}}e^{i\mathbf{k}\,\cdot\,\mathbf{x}} \qquad \text{and} \qquad \psi_{\mathbf{k}}^*(\mathbf{x}) = \frac{1}{\sqrt{V}}e^{-i\mathbf{k}\,\cdot\,\mathbf{x}}, \tag{3.103}$$

where we have $\psi_{\mathbf{k}}^* = \psi_{-\mathbf{k}}$ and $\nabla_{\mathbf{x}}^2\psi_{\pm\mathbf{k}}(\mathbf{x}) = \mathbf{k}^2\psi_{\pm\mathbf{k}}(\mathbf{x})$. To each pair of spatial solutions with wave vector $\pm\mathbf{k}$ corresponds a pair of linearly independent temporal solutions $h_k(t)$, whose quadratic frequencies $\Omega_k^2(t)$ are given by (see section 4.2):

$$\Omega_k^2(t) = \omega_k^2(t) + \sigma(t), \tag{3.104}$$

$$\omega_k(t) = \sqrt{\frac{k^2}{a^2(t)} + m^2}, \qquad \sigma(t) = \left(6\xi - \frac{3}{4}\right)\frac{\dot{a}^2}{a^2} + \left(6\xi - \frac{3}{2}\right)\frac{\ddot{a}}{a}.$$

Then, *in the asymptotic regions*, these pairs of *exact solutions* can be decomposed in positive and negative frequency solutions $\{h_k^{(p)}, h_k^{*(p)}\}$ and $\{h_k^{(f)}, h_k^{*(f)}\}$, such that:

$$\begin{cases} i\dfrac{d}{dt}h_k^{(p)}(t) \simeq \omega_{k1}h_k^{(p)}(t), \qquad i\dfrac{d}{dt}h_k^{*(p)}(t) \simeq -\omega_{k1}h_k^{*(p)}(t), \qquad x \in \Omega_p \ (t \to -\infty) & (3.105a) \\ i\dfrac{d}{dt}h_k^{(f)}(t) \simeq \omega_{k2}h_k^{(f)}(t), \qquad i\dfrac{d}{dt}h_k^{*(f)}(t) \simeq -\omega_{k2}h_k^{*(f)}(t), \qquad x \in \Omega_f \ (t \to +\infty) & (3.105b) \end{cases}$$

where we have defined the past and future asymptotic frequencies:

$$\Omega_{k(1,2)} = \omega_{k(1,2)} \equiv \sqrt{\frac{k^2}{a_{(1,2)}^2} + m^2}. \tag{3.106}$$

Here, we stress once again that *solutions belonging to different* $\{\mathbf{k}, -\mathbf{k}\}$ *pairs remain orthogonal at all times*. Thus, to evaluate particle creation, we just have to consider (2x2) block diagonal Bogolubov transformations among the pairs $\{u_{\mathbf{k}}^{(p)}, u_{\mathbf{k}}^{*(p)}\}$ and $\{u_{\mathbf{k}}^{(f)}, u_{\mathbf{k}}^{*(f)}\}$, from which we find:

$$\begin{cases} \alpha_{\mathbf{k}\mathbf{k}'} = \alpha_k \delta_{\mathbf{k},\mathbf{k}'}, & (3.107a) \\ \beta_{\mathbf{k}\mathbf{k}'} = \beta_k \delta_{\mathbf{k},-\mathbf{k}'}, & (3.107b) \end{cases}$$

where (i) the coefficients α_k and β_k only depend on the magnitude of $\mathbf{k}$ as a consequence of spatial isotropy and (ii) we have enforced that the α coefficients must be strictly diagonal $(\mathbf{k}, \mathbf{k})$, whereas the β ones must be crossed $(\mathbf{k}, -\mathbf{k})$, since the spatial $(\mathbf{x})$ dependence for $u_{\mathbf{k}}$ $(u_{\mathbf{k}}^*)$ is given by $\psi_{\mathbf{k}}$ $(\psi_{\mathbf{k}}^*)$ *at all times*. More explicitly:

$$\begin{aligned} u_{\mathbf{k}}^{(f)}(x) = h_k^{(f)}(t)\psi_{\mathbf{k}}(\mathbf{x}) &= \alpha_k u_{\mathbf{k}}^{(p)}(x) + \beta_k u_{-\mathbf{k}}^{*(p)}(x) \\ &= \Big(\alpha_k h_k^{(p)}(t) + \beta_k h_k^{*(p)}(t)\Big)\psi_{\mathbf{k}}(\mathbf{x}) \end{aligned} \tag{3.108}$$

(had we had contributions from $u_{-\mathbf{k}}^{(f)}$ or $u_{\mathbf{k}}^{*(f)}$, we would end up with terms proportional to $\psi_{\mathbf{k}}^* = \psi_{-\mathbf{k}}$, and the equality with the LHS could not match).

In the special case (3.107), there are great simplifications in the relations between both modes and results for particle creation. For example, the expected value (3.53) for the total number of particles measured in the asymptotic future, starting from a vacuum state in the past, will be just:

$$\langle 0_p | N^{(f)} | 0_p \rangle = \sum_{\mathbf{k}} |\beta_k|^2. \tag{3.109}$$

Further, the consistency condition (3.47) for the Bogolubov coefficients greatly simplify to:

$$|\alpha_k|^2 - |\beta_k|^2 = 1. \tag{3.110}$$

As in the general case, these will be compatible with commutation relations, while anticommutation relations would yield (3.50):

$$|\alpha_k|^2 + |\beta_k|^2 = 1 \tag{3.111}$$

which are only compatible with (3.110) when *all* β_k's are null, that is, when there are no created particles whatsoever. Thus, in this particular context, where the Bogolubov coefficients can be interpreted dynamically in terms of particle creation, one could argue (as in (2)) *that the scalar (spin 0) field statistics must be bosonic in curved spacetimes by virtue of its dynamics.* We stress that, generally, a bosonic statistic is enforced *as a consistency condition* (so that one may perform the quantization on equal footing for any orthonormal mode expansion), whether or not one may interpret it dynamically. Nonetheless, it is interesting that in some special contexts, one can make such dynamical interpretation of the spin-statistics relation.

Finally, we note that the mode operators can be written in terms of one another as (eqs 3.42):

$$a^{(f)}_{\mathbf{k}} = \alpha_k a^{(p)}_{\mathbf{k}} + \beta^*_k \, a^{\dagger(p)}_{-\mathbf{k}}, \tag{3.112a}$$
$$a^{(p)}_{\mathbf{k}} = \alpha^*_k a^{(f)}_{\mathbf{k}} - \beta^*_k \, a^{\dagger(f)}_{-\mathbf{k}}. \tag{3.112b}$$

Then, since these transformations are 'quasidiagonal', they are extremely simpler to invert than in the general case. These will allow us to compute vacuum to many-particle state projections with considerable ease, which we shall use to analyze the statistics and correlations for created particles in the next subsection.

3.4.2 Correlations and statistics of created particles

We have seen that asymptotically flat FLRW spaces make a very convenient stage to analyze particle creation, and so far we have found that one can find the total expectation values for particles in the asymptotic future by (3.109) (or for particles of each type, $\langle N^{(f)}_{\mathbf{k}}\rangle$, by withholding the sum and just looking at a particular $\mathbf{k}$ value). However, these expectation values alone do not tell us all about the statistics of the created particles; they just convey information about its *averages*. Indeed, it is easy to see that, for instance, the states $|\Psi_1\rangle = |1_{\mathbf{k}}\rangle$ and $|\Psi_2\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |2_{\mathbf{k}}\rangle)$ both yield the same expected values:

$$\langle\Psi_1|N_{\mathbf{k}'}|\Psi_1\rangle = \langle\Psi_2|N_{\mathbf{k}'}|\Psi_2\rangle = 1 \times \delta_{\mathbf{k},\mathbf{k}'}, \tag{3.113}$$

even though $\langle\Psi_1|\Psi_2\rangle = 0$. To obtain a more detailed statistical information of the created particles, we must analyze general transition amplitudes of the form $\langle n^{(f)}_{\mathbf{k_1}}, n^{(f)}_{\mathbf{k_2}}, ...|0_p\rangle$. Before we analyze these in full generality, it is constructive that we look at more simple particle states. The fact that the relation (3.112) mixes $\mathbf{k}$ and $-\mathbf{k}$ modes is suggestive that it will be useful to start with amplitudes of the form:

$$\mathcal{A}_n(\mathbf{k}) \equiv \langle n^{(f)}_{\mathbf{k}}, n^{(f)}_{-\mathbf{k}}, |0_p\rangle\,, \tag{3.114}$$

i.e. the probability amplitude that n pairs of particles were created in the modes $u^{(f)}_{\mathbf{k}}$ and $u^{(f)}_{-\mathbf{k}}$ (and no others). We have that:

$$\begin{aligned}
\mathcal{A}_n(\mathbf{k}) &= \frac{1}{n!}\,\langle 0_f|(a^{(f)}_{\mathbf{k}})^n(a^{(f)}_{-\mathbf{k}})^n|0_p\rangle \\
&= \frac{1}{n!}\,\langle 0_f|(a^{(f)}_{\mathbf{k}})^n(\alpha_k a^{(p)}_{-\mathbf{k}} + \beta^*_k a^{\dagger(p)}_{\mathbf{k}})^n|0_p\rangle \\
&= \frac{1}{n!}(\beta^*_k)^n\,\langle 0_f|(a^{(f)}_{\mathbf{k}})^n(a^{\dagger(p)}_{\mathbf{k}})^n|0_p\rangle \\
&= \frac{1}{n!}(\beta^*_k)^n\,\langle 0_f\Big|(a^{(f)}_{\mathbf{k}})^n\left(\frac{a^{\dagger(f)}_{\mathbf{k}} - \beta_k a^{(p)}_{-\mathbf{k}}}{\alpha^*_k}\right)^n\Big|0_p\rangle \\
&= \frac{1}{n!}\left(\frac{\beta^*_k}{\alpha^*_k}\right)^n\,\langle 0_f\Big|(a^{(f)}_{\mathbf{k}})^n(a^{\dagger(f)}_{\mathbf{k}})^n\Big|0_p\rangle \\
&= \frac{1}{(n-1)!}\left(\frac{\beta^*_k}{\alpha^*_k}\right)^n\,\langle 0_f\Big|(a^{(f)}_{\mathbf{k}})^{n-1}(a^{\dagger(f)}_{\mathbf{k}})^{n-1}\Big|0_p\rangle \\
&\quad\vdots \\
&= \left(\frac{\beta^*_k}{\alpha^*_k}\right)^n\,\langle 0_f|0_p\rangle\,,
\end{aligned} \tag{3.115}$$

where, in the last lines, we have recursively applied the commutation relations $[(a_{\mathbf{k}})^n, a^\dagger_{\mathbf{k}}] = n(a_{\mathbf{k}})^{n-1}$. From these same lines it is also easy to see that, for $m \neq n$:

$$\langle m^{(f)}_{\mathbf{k}}, n^{(f)}_{-\mathbf{k}}, |0_p\rangle = 0. \tag{3.116}$$

Therefore, we conclude that particles are always produced in pairs with the same energy and opposite momenta. Indeed, this is to be expected in FLRW spacetimes, since spatial homogeneity implies the conservation of 3-momentum. Note, however, that we have deduced a stronger restriction, since conservation of momentum alone could still allow for created particles in sets like $|1_{\mathbf{k}}, 2_{-\mathbf{k}/2}\rangle$, $|1_{\mathbf{k}}, 3_{-\mathbf{k}/3}\rangle$ and other similar combinations. The restriction we have just deduced means that *the only states in the Fock Space built on $|0_f\rangle$ that are not orthogonal to $|0_p\rangle$ are those built with pairs of particles in the modes $u^{(f)}_{\mathbf{k}}$ and $u^{(f)}_{-\mathbf{k}}$*. For brevity, we drop the superscripts (f) of the future modes and denote these states as:

$$|\{n_j(\mathbf{k}_j)\}\rangle = |{}^1n_{\mathbf{k_1}}, {}^1n_{-\mathbf{k_1}}; {}^2n_{\mathbf{k_2}}, {}^2n_{-\mathbf{k_2}}; ...\rangle\,. \tag{3.117}$$

We can then write a completeness relation for $|0_p\rangle$:

$$|0_p\rangle = \sum_{\{n_j(\mathbf{k}_j)\}} |\{n_j(\mathbf{k}_j)\}\rangle\langle\{n_j(\mathbf{k}_j)\}|0_p\rangle = \sum_{\{n_j(\mathbf{k}_j)\}} |\{n_j(\mathbf{k}_j)\}\rangle \left[\prod_j \left(\frac{\beta_k^*}{\alpha_k^*}\right)^{n_j}\right] \langle 0_f|0_p\rangle \, . \tag{3.118}$$

From this equation, we can compute the norm of the vacuum to vacuum transition $|\langle 0_f|0_p\rangle|$ using the normalization condition:

$$1 = |\langle 0_p|0_p\rangle|^2 = \left(\sum_{\{n_j(\mathbf{k}_j)\}} \prod_j \left|\frac{\beta_k}{\alpha_k}\right|^{2n_j}\right) |\langle 0_f|0_p\rangle|^2. \tag{3.119}$$

Well, assuming all summations and products converge appropriately, we may commute them, by noting that:

$$\begin{aligned}\sum_{\{n_j\}} \prod_j x_j^{n_j} &= \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} \ldots (x_1)^{n_1}(x_2)^{n_2}\ldots \\ &= \Big(\sum_{n_1=0}^{\infty} x_1^{n_1}\Big)\Big(\sum_{n_2=0}^{\infty} x_2^{n_2}\Big)\ldots \\ &= \prod_j \Big(\sum_{n_j=0}^{\infty} x_j^{n_j}\Big), \end{aligned} \tag{3.120}$$

where these summations are just familiar geometric series. Thus, we have:

$$\begin{aligned} 1 &= |\langle 0_f|0_p\rangle|^2 \prod_j \left(\sum_{n_j=0}^{\infty} \left|\frac{\beta_{k_j}}{\alpha_{k_j}}\right|^{2n_j}\right) \\ &= |\langle 0_f|0_p\rangle|^2 \prod_j \left[1 - \left|\frac{\beta_{k_j}}{\alpha_{k_j}}\right|^2\right]^{-1} \\ &= |\langle 0_f|0_p\rangle|^2 \prod_j |\alpha_{k_j}|^2, \end{aligned} \tag{3.121}$$

where we have used (3.110). Eq (3.121) immediately entails:

$$|\langle 0_f|0_p\rangle|^2 = \prod_j |\alpha_{k_j}|^{-2}. \tag{3.122}$$

Finally, we obtain the explicit transition probabilities:

$$P\Big(\{n_j(\mathbf{k}_j)\}\Big) \equiv \Big|\langle\{n_j(\mathbf{k}_j)\}|0_p\rangle\Big|^2 = \left[\prod_j |\alpha_{k_j}|^{-2}\left|\frac{\beta_{k_j}}{\alpha_{k_j}}\right|^{2n_j}\right]. \tag{3.123}$$

Using this expression, it is particularly interesting to note the marginal probabilities that emerge for the creation of n pairs of just one type. If we fix only one of the n_j (setting $n_j = n$) in (3.123), and sum over all the possibilities for the remaining modes (with $j' \neq j$), we obtain the probability that n pairs will be created in mode $\mathbf{k}$:

$$P_n(\mathbf{k}) \equiv \Big|\mathcal{A}_n(\mathbf{k})\Big|^2 = |\alpha_k|^{-2}\left|\frac{\beta_k}{\alpha_k}\right|^{2n}. \tag{3.124}$$

We could also compute a marginal probability for creating n_1 particles in a mode $\mathbf{k_1}$ *and* n_2 particles in a mode $\mathbf{k_2}$. From eq. (3.123), we see immediately that pair production for distinct modes ($j \neq j'$) are independent events, since their joint probability is just the product of the individual marginal probabilities:

$$P\Big(n_1(\mathbf{k}_1), n_2(\mathbf{k}_2)\Big) = P_{n_1}(\mathbf{k}_1)P_{n_2}(\mathbf{k}_2). \tag{3.125}$$

However, *the production of multiple pairs in the same mode* $\mathbf{k}$ *are not independent events*. One can see directly from (3.124) that:

$$P_2(\mathbf{k}) = \frac{P_1^2(\mathbf{k})}{P_0(\mathbf{k})} \geq P_1^2(\mathbf{k}), \tag{3.126}$$

where the equality will only occur for $\beta_k = 0$, when particle creation in mode $\mathbf{k}$ is trivial ($P_0(\mathbf{k}) = 1$, $P_{n\geq 1}(\mathbf{k}) = 0$). More generally, we have:

$$P_n(\mathbf{k})\frac{P_1^n(\mathbf{k})}{P_0^{n-1}(\mathbf{k})} = \left(\frac{P_1(\mathbf{k})}{P_0(\mathbf{k})}\right)^n P_0(\mathbf{k}) \geq P_1^n(\mathbf{k}). \tag{3.127}$$

Thus, the probability of creating n pairs in the same mode $\mathbf{k}$ is generally greater than the probability of creating all of these pairs independently. This is analogous to the phenomena of spontaneous and stimulated emission (*e.g.* for atoms interacting with radiation), where the probability of emitting one more photon increases as there are more photons present.

From eq. (3.124) it is easy to recover the known *average* results for particle creation. In fact, it is not hard to compute any statistical moments; in zeroth order, we reobtain the normalization of probability:

$$\sum_{n=0}^{\infty} P_n(\mathbf{k}) = 1. \tag{3.128}$$

Here one must just sum geometric series, as in (3.121). Then, the first order moment recovers the average/expected value:

$$\begin{aligned}\langle 0_p|N_{\mathbf{k}}|0_p\rangle &= \sum_{n=0}^{\infty} \langle 0_p|n_{\mathbf{k}}, n_{-\mathbf{k}}\rangle \langle n_{\mathbf{k}}, n_{-\mathbf{k}}|a_{\mathbf{k}}^{\dagger}a_{\mathbf{k}}|n_{\mathbf{k}}, n_{-\mathbf{k}}\rangle \langle n_{\mathbf{k}}, n_{-\mathbf{k}}|0_p\rangle \\ &= \sum_{n=0}^{\infty} \langle 0_p|0_f\rangle \left(\frac{\beta_k}{\alpha_k}\right)^n n \left(\frac{\beta_k^*}{\alpha_k^*}\right)^n \langle 0_f|0_p\rangle \\ &= \sum_{n=0}^{\infty} nP_n(\mathbf{k}) \\ &= |\beta_k|^2, \end{aligned} \tag{3.129}$$

where we already know the last equality to be true from the Bogolubov transformations (3.109). Still, it is not difficult to compute it directly through the summation $\sum_n nP_n$ by employing a little trick of taking partial derivatives with respect to β_k:

$$\begin{aligned}\sum_{n=0}^{\infty} nP_n(\mathbf{k}) &= \sum_{n=0}^{\infty} n\left|\frac{\beta_k}{\alpha_k}\right|^{2n} |\alpha_k|^{-2} \\ &= |\beta_k|^2 \frac{\partial}{\partial|\beta_k|^2} \sum_{n=0}^{\infty}\left|\frac{\beta_k}{\alpha_k}\right|^{2n} |\alpha_k|^{-2} \\ &= \left|\frac{\beta_k}{\alpha_k}\right|^2 \frac{\partial}{\partial|\beta_k|^2}\left(1-\left|\frac{\beta_k}{\alpha_k}\right|^2\right)^{-1} \\ &= \left|\frac{\beta_k}{\alpha_k}\right|^2 |\alpha_k|^{-2}\left(1-\left|\frac{\beta_k}{\alpha_k}\right|^2\right)^{-2} \\ &= |\beta_k|^2. \end{aligned} \tag{3.130}$$

(In implementing this trick, however, one must be careful to *only impose eq* (3.110) *after taking the derivatives with respect to* $|\beta_k|$, as treated α_k and β_k as independent variables to write the second equality.)

Then, if one wishes, it is possible to carry analogous calculations for higher statistical moments (such as the variance).

With the above results, we can recover the expected value for the total particle density in the asymptotic future, due created particles. Particularly, taking the continuum limit, we obtain:

$$\begin{aligned}\langle 0_p|N|0_p\rangle &= \lim_{L\to\infty} \frac{1}{La_2^3}\sum_{\mathbf{k}} |\beta_k|^2 \\ &= \frac{1}{2\pi^2 a_2^3}\int_0^{\infty} dk\, k^2 |\beta(k)|^2 \end{aligned}$$

$$= \frac{1}{2\pi^2} \int_0^\infty dk'(k')^2 |\beta(a_2 k')|^2, \tag{3.131}$$

where we have absorbed the scale factor a_2 in the definition of the physical momentum in the asymptotic future $k' \equiv k/a_2$.

3.4.3 A simple model for particle creation

Now that we have developed many features of particle creation in a model-independent way[X], we would like to better grasp this phenomenon through a simple, tractable model, for which we can explicitly compute the Bogolubov coefficients. This shall serve both to illustrate the general (dynamic-independent) features presented so far, and to give a glimpse of how particle creation ultimately depends on the *dynamics* of spacetime in its nonstationary phase, preparing the ground for how we may define suitable extensions of (approximate) concepts of vacuum and particles *to fully dynamical spacetimes* (letting go the hyphothesis of asymptotic flatness). Here, we shall explore a simple model presented in section 3.4 of (1).

For simplicity, this model is built in $1+1$ spacetime dimensions in a FLRW metric. Here we make explicit use of the conformally flat form of the metric (see appendix B), writing it in conformal coordinates (η, x):

$$ds^2 = dt^2 - a^2(t)dx^2 = a^2(\eta)\Big[d\eta^2 - dx^2\Big], \tag{3.132}$$

where η is called the *conformal time*, defined by: $\eta = \int^t \frac{dt'}{a(t')}$. We then define the scale factor as a function of η to be (see Figure 18) :

$$a^2(\eta) = A + B\tanh(\rho\eta), \tag{3.133}$$

where A, B and ρ are constant parameters. Note that $a^2(\eta) \to A \pm B$ as $\eta \to \pm\infty$.

[X] That is, we have not assumed a particular metric. Even when we specialized to FLRW metrics, we have not assumed a specific form for $a(t)$.

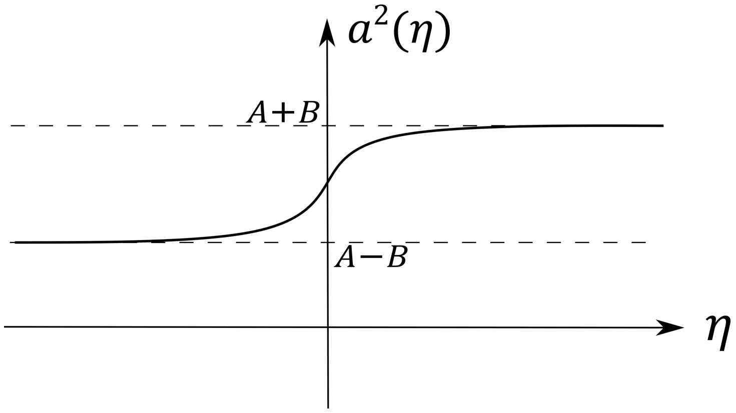

Figure 18 – Scale factor for a simple model of expansion displayed as a function of conformal time in the asymptotic regions $\eta \to \pm\infty$ it becomes asymptotically flat, with $a \to (A \pm B)$.
Source: By the author.

We leave to the next section a more thorough discussion of the form and solutions to the massive field equations for a conformally flat spacetime. For the time being, we note that, analogously to when we employed proper-time coordinates, we shall obtain separable solutions, with simple exponential dependences in space, and time-dependent harmonic oscillators in time, whose frequencies $\omega_k(\eta)$ are given by (3.149). This model then yields the asymptotic frequencies for each wave vector $\mathbf{k}$ (we omit the k subscript for cleaness) in the far past and future:

$$\begin{cases} \omega_1 \equiv \sqrt{k^2 + m^2(A-B)} & (3.134a) \\ \omega_2 \equiv \sqrt{k^2 + m^2(A+B)} & (3.134b) \end{cases} .$$

For later convenience, we also define the frequencies:

$$\omega_\pm \equiv \frac{1}{2}\Big(\omega_2 \pm \omega_1\Big). \tag{3.135}$$

The exact field equations will be given by (3.148) (with the scale factor (3.133)), for which it is possible to obtain the (normalized) exact mode solutions $u_{\mathbf{k}}^{(p)}$ and $u_{\mathbf{k}}^{(f)}$, which behave as positive frequencies in the asymptotic past and the asymptotic future, respectively. They read (see (1)):

$$\begin{aligned} u_k^{(p)}(\eta, x) &= \frac{1}{\sqrt{4\pi\omega_1}} e^{ikx - i\omega_+\eta - i\frac{\omega_-}{\rho}\ln[2\cosh(\rho\eta)]} \times {}_2F_1\Big(1 + i\tfrac{\omega_-}{\rho}, i\tfrac{\omega_-}{\rho}; 1 - i\tfrac{\omega_1}{\rho}; \tfrac{1}{2}(1 + \tanh\rho\eta)\Big) \\ &\longrightarrow \frac{1}{\sqrt{4\pi\omega_1}} e^{ikx - i\omega_1\eta}, \quad \text{as } \eta \to -\infty, \qquad (3.136) \\ u_k^{(f)}(\eta, x) &= \frac{1}{\sqrt{4\pi\omega_2}} e^{ikx - i\omega_+\eta - i\frac{\omega_-}{\rho}\ln[2\cosh(\rho\eta)]} \times {}_2F_1\Big(1 + i\tfrac{\omega_-}{\rho}, i\tfrac{\omega_-}{\rho}; 1 - i\tfrac{\omega_2}{\rho}; \tfrac{1}{2}(1 - \tanh\rho\eta)\Big) \end{aligned}$$

$$\longrightarrow \frac{1}{\sqrt{4\pi\omega_2}} e^{ikx - i\omega_2\eta}, \quad \text{as } \eta \to +\infty, \tag{3.137}$$

where ${}_2F_1$ is a hypergeometric function. Here, we need not to worry about the details in obtaining these solutions; it is not difficult to verify those indeed satisfy the field equations and have the appropriate asymptotic limits (see section 9.1 of (21)). Also, it is easy to see that those solutions do not coincide, such that one will generally have nonzero β_k coefficients and there will be particle creation. One may verify (see (1) and references therein, or section 7.5 of (21)) that the Bogolubov transformations take the form:

$$u_k^{(p)}(\eta, x) = \alpha_k u_k^{(f)}(\eta, x) + \beta_k u_{-k}^{(f)}(\eta, x), \tag{3.138}$$

with:

$$\begin{cases} \alpha_k = \left(\dfrac{\omega_2}{\omega_1}\right)^{1/2} \dfrac{\Gamma(1 - i\omega_1/\rho)\Gamma(-i\omega_2/\rho)}{\Gamma(1 - i\omega_+/\rho)\Gamma(-i\omega_+/\rho)} & \text{(3.139a)} \\[2ex] \beta_k = \left(\dfrac{\omega_2}{\omega_1}\right)^{1/2} \dfrac{\Gamma(1 - i\omega_1/\rho)\Gamma(i\omega_2/\rho)}{\Gamma(1 + i\omega_-/\rho)\Gamma(i\omega_-/\rho)} & \text{(3.139b)} \end{cases} .$$

Then, using the properties of the gamma function:

$$\begin{aligned} \Gamma(1 + x) &= x\Gamma(x), \\ |\Gamma(iy)|^2 &= \frac{\pi}{y \sinh(\pi y)}, \end{aligned}$$

one can immediately obtain the quadratic Bogolubov coefficients:

$$\begin{cases} |\alpha_k|^2 = \dfrac{\sinh^2(\pi\omega_+/\rho)}{\sinh(\pi\omega_1/\rho)\sinh(\pi\omega_2/\rho)} & \text{(3.140a)} \\[2ex] |\beta_k|^2 = \dfrac{\sinh^2(\pi\omega_-/\rho)}{\sinh(\pi\omega_1/\rho)\sinh(\pi\omega_2/\rho)} & \text{(3.140b)} \end{cases} .$$

In this form, it is easy to verify the Bogolubov condition:

$$|\alpha_k|^2 - |\beta_k|^2 = 1, \tag{3.141}$$

and to verify a few consistency checks. For example, this formula gives us a quite intuitive dependence on the frequencies; particularly, the $|\beta_k|^2$, which accounts from particle creation is proportional to $\sinh^2(\pi\omega_-/\rho)$, so that it increases the more ω_2 differs from ω_1, and vanishes in the limit of no expansion ($B \to 0$), when $\omega_2 \to \omega_1$. Note that this will always be the case for a conformally coupled massless field, whose frequencies remain in the form $\omega_k = k$. One may then interpret that the mass, which breaks conformal invariance, couples the field nontrivially to gravity, allowing the spacetime expansion to inject it with the energy necessary for particle creation. Furthermore, even in the massive case, note that this frequency difference becomes progressively smaller for higher values of k, such that creation of particles will be suppressed for arbitrary high-energy, short-wavelength modes. We shall discuss these features in more detail in the next section, where we will try to circumscribe an appropriate extension to the concepts of vacuum and particles in more general dynamic spacetimes.

3.5 Adiabatic vacuum

As we have seen in the last section, dynamical spacetimes will generally not possess a distinguished notion of vacuum, even if we restrict ourselves to inertial (free falling) observers. Particularly, when there were asymptotically flat regions of our spacetime, this phenomenon could be better grasped in terms of particle creation, which could be analyzed simply in terms of asymptotically positive-frequency modes in the far past and future.

In such context, given that there are particles present *after* the expansion, but not *before* (as measured by any inertial observers in the asymptotic regions Ω_f and Ω_p, respectively), one may be tempted to infer that the particle creation must have ocurred *during* the expansion, and, thus, that measurements performed between these regions would yield an intermediate number of particles. However, these claims do not survive upon closer inspection. As we have thoroughly discussed in section 3.3, detectors in a dynamical region will generally respond in a quite complicated way to their interactions with the field, and one should not *a priori* expect them to measure an intermediate particle content between Ω_p and Ω_f.

As discussed above, there are no physically privileged definitions of vacuum and particles for general spacetimes, so that it is not always possible to define particle number without ambiguity away from asymptotically flat regions. We have also seen that a special class of *dynamical* spacetimes for which a privileged (nonstationary) family of inertial observers does exist, are the FLRW spaces, with their comoving observers. In this case, we could try to identify the presence of particles *throughout dynamical regions* according to the detection rates for particle detectors carried by these observers.

However, even in such highly symmetric cases, for which a preferred physical definition of particles is possible, particle numbers are not conserved quantities in nonstationary spaces, which makes their measurement inherently uncertain. If, say, the rate of particle creation is A, then a precise measurement of particle numbers in a given time must be carried in a sufficiently short time window Δt, such that $|A|\Delta t \ll 1$. However there is a fundamental limitation on how short Δt can be not to violate the (time-energy) uncertainty principle[Y] : if one is to make a detection of an excitation within a precision ΔE, we must have that $\Delta t \gtrsim \Delta E^{-1}$. Since any single particle will cause an excitation of *at least* $\Delta E_1 = m$, a precise detection of N particles will be associated with a minimal time interval $\Delta t \sim \Delta E_N^{-1} \sim (m\Delta N)^{-1}$. Taking into account both sources of uncertainty, we have a rough estimate on the limits on the precision for measuring N:

$$\Delta N \gtrsim (m\Delta t)^{-1} + |A|\Delta t, \tag{3.142}$$

so that we have a minimal uncertainty $\Delta N_{min} \sim 2(|A|/m)^{1/2}$ for $\Delta t = (m|A|)^{-1/2}$.

Thus we see that, for a nonzero particle creation rate $A \neq 0$, and a field of finite mass, there is a fundamental limitation in the precision of particle measurements for any modes in a given time[Z]. Nevertheless, note that the number uncertainty for arbitrarily high-energy (short-wavelenghth) particles – for which $\Delta E_1 \simeq \sqrt{\mathbf{k}^2 + m^2} \gg m$ – will be vanishingly small. This will be true even if the particle creation rate happens to be the same for all modes, since the higher energy modes are associated with smaller time uncertainty in measurements. But besides that, particle production itself is generally suppressed at high energies, as we have seen in last section.

Having pointed out these fundamental limitations, we know nonetheless that there must be appropriate limits for which particle numbers must be meaningful observables. Particularly, given the astounding success of QFT in Minkowski spacetime to describe our terrestrial experiments, as well as high energy astrophysical observations, one should expect to reobtain this theory as a sufficiently good approximation for QFTCS in our own expanding universe. Furthermore, this approximation should be increasingly better for a correspondingly slower rate of expansion.

To investigate these considerations more concretely, we turn to the simple model that was presented at the end of the previous section. There, we found that particle creation was supressed when $\omega_- \to 0$, and that would indeed occur in the limit of no expansion; more precisely, if we take $B \to 0$ and expand $|\beta_k|^2$ to the lowest order in B, we find that it decays as $|\beta_k|^2 \propto B^2 \to 0$.

Moreover, upon closer inspection of (3.140b), we find that, for a fixed value of *total expansion* B, if we take the expansion *rate* ρ to vanishingly small values, particle creation will be exponentially suppressed for all modes, *i.e.*:

[Y] For a thorough and pedagocical exposition of time-energy uncertainty principle, see chapter 3 of (30) and references therein.

[Z] Note that, in the massless case, this uncertainty diverges. This is due to the fact that one can have particles of arbitrarily low energies, which would require arbitrarily high precision to be accounted for. This problem is part of a more general hall of infrared divergencies that occur for massless theories, even in Minkowski space. For an introductory account of these divergencies in the simple context of Minkowski spacetimes, see chapter 7 of (6).

$$|\beta_k|^2 \propto e^{-2\pi\omega_p/\rho} \to 0, \qquad \rho \ll |\mathbf{k}|, m\sqrt{B}. \tag{3.143}$$

For any finite ρ, this exponential suppression will hold approximately for $\rho \ll \omega_p$, that is, always that $\rho \ll k$ or $\rho \ll B^{\frac{1}{2}}m$. Physically, we can interpret this suppression as a limitation in the production of particles for modes whose frequencies are much larger than the relative rate of expansion $\omega \gg \dot{a}/a \equiv H$ of the universe, so that particle creation should be negligible for high-energy modes of all fields (and, particularly, for any modes of a very massive field); we only expect there to be an appreciable amount of created particles for modes of frequencies $\omega \lesssim H$, comparable to the fractional rate of expansion of the universe or lower. (For our current universe, we see that this rate is extremely low: $H_0 \approx 2 \times 10^{-18}s$, corresponding to energies of about $\hbar H_0 \approx 8 \times 10^{-33}eV$.)

Although we have only deduced these conditions in the context of a simple model, they turn out to be valid in general (1). Then, to generalize an approximate notion of particles *during* the expansion, it will be particularly useful to refer to the limits of very slow expansions, which will allow us to construct a corresponding approximation for positive-frequency modes. Such an approximation should become increasingly precise as the rate of expansion becomes arbitrarily slower; in the limit of an infinitely slow expansion – which we will baptize as the *adiabatic limit* ahead –, they should be exact, matching the fact that there will be no particles created.

As in the discussion of particle creation, it may be convenient to work with either conformal time or proper time, depending on the application at hand. The former is somewhat easier to handle the dynamical equations (especially in the massless, conformally invariant limit), as well as to present adiabatic expansions and discuss adiabatic orders in an algebraic manner, and we shall employ it in the present section. The latter, on its turn, has a more direct physical interpretation and can be worked without much difficulty with a conveniently chosen decomposition of field modes; we shall discuss it in more detail in the next chapter, in the context of adiabatic subtractions.

In order to keep the discussion simple in this first exposition, we restrict the treatment in this section to FLRW spacetimes conformally related to Minkowski spacetime, and to fields conformally coupled to gravity, *i.e.* with $\xi = \xi(n)$ in n spacetime dimensions (see Apendix B.3). In this case, we write the line element as:

$$ds^2 = a^2(\eta)[d\eta^2 - d\mathbf{x}^2]. \tag{3.144}$$

Implementing the conformal transformation to the field equations in this case (again, see appendix B.3), and accounting for the presence of a mass term, which adds a conformally noninvariant contribution, but which can be accounted for with a simple term $m^2\tilde{\phi} = m^2\Omega^s\phi$ in the transformed equation, we obtain:

$$\Big[\tilde{\Box}_x + \xi(n)\tilde{R}(x) + m^2\Big]\tilde{\phi}(x) = \Omega^{s-2}\Big[\Box_x + \xi(n)R(x) + m^2\Omega^2(x)\Big]\phi(x) = 0. \tag{3.145}$$

For a conformally flat metric $\tilde{g}_{ab} = \Omega^2\eta_{ab}$, we have simply $R(x) = 0$ and $\Box_x = \frac{\partial^2}{\partial\eta^2} - \nabla^2_{\mathbf{x}}$. Furthermore, the conformal factor Ω is seen to be simply the scale factor of the universe, $\Omega^2(\eta, \mathbf{x}) = a^2(\eta)$, so that we can simplify the right equation in (3.145) to:

$$\Big[\frac{\partial^2}{\partial\eta^2} - \nabla^2_{\mathbf{x}} + m^2a^2(\eta)\Big]\phi(x) = 0. \tag{3.146}$$

This is a manifestly separable equation, so that we can decompose normal modes $v_{\mathbf{k}}$ (we save the notation $u_{\mathbf{k}}$ for the solutions of the original field equation $u_k = \Omega^{s-2}v_{\mathbf{k}}$) in the form:

$$v_{\mathbf{k}}(\eta, \mathbf{x}) = \frac{e^{i\mathbf{k}\,\cdot\,\mathbf{x}}}{(2\pi)^{3/2}}\chi_k(\eta). \tag{3.147}$$

Substituting them in (3.145), we find that χ_k will be just a time-dependent harmonic oscillator:

$$\frac{d^2}{d\eta^2}\chi_k + \omega_k^2(\eta)\chi_k = 0, \tag{3.148}$$

where the positive frequencies $\omega_k(\eta)$ are defined by:

$$\omega_k(\eta) \equiv +\sqrt{\mathbf{k}^2 + m^2a^2(\eta)}. \tag{3.149}$$

Once again, in the case of static universe $a = cte$, we trivially recover plane-wave modes, $\chi_k = (2\omega_k)^{-1/2}e^{-i\omega_k\eta}$. In the general case, however, not only are these equations hard to solve, but also their solution space cannot be globally separated in positive and negative frequency subspaces. To try to make sense of such a separation *locally*, we write formal WKB solutions, which take the form:

$$\chi_k = \frac{1}{\sqrt{W_k(\eta)}}e^{-i\int^{\eta}d\eta' W_k(\eta')}. \tag{3.150}$$

Substituting these in (3.148), we obtain a nonlinear equation for W_k:

$$\begin{aligned} W_k^2(\eta) &= \omega_k^2(\eta) - \frac{1}{2}\Big(\frac{\ddot{W}_k}{W_k} - \frac{3}{2}\frac{\dot{W}_k^2}{W_k^2}\Big) \\ &= \omega_k^2(\eta) + W_k^{\frac{1}{2}}\frac{d^2}{d\eta^2}W_k^{-\frac{1}{2}}. \end{aligned} \tag{3.151}$$

At this point, the reader may wonder why one would choose to work with this rather complicated nonlinear equation (3.151) instead of (3.148) directly. The advantage here lies not in obtaining exact field solutions to these equations, but rather in analyzing their behaviour in the limit of a very slow expansion, when the time derivatives of W_k (ω_k) become negligible. Particularly in the limit of an infinitely slow expansion, we will have that $\dot{\omega}_k \to 0$, and (3.151) will yield a purely algebraic relation:

$$W_k(\eta) = \omega_k(\eta), \tag{3.152}$$

where we have identified the positive roots of W_k and ω_k. For a finitely slow expansion – where it still holds that $\dot{\omega} \ll \omega^2$, as well as the inequalities obtained through further time derivatives: $\ddot{\omega} \ll 2\omega\dot{\omega} \ll 2\omega^3$, etc. –, equation (3.152) can be regarded as a zeroth order approximation for W_k.

In order to refine that approximation beyond lowest order in a systematic manner, and quantify a more precise notion of 'slowness', we introduce the so-called *adiabatic parameter* T, transforming the time variable as $\eta \to \eta_1 \equiv \eta/T$:

$$\eta \to \eta_1 \equiv \frac{\eta}{T}; \tag{3.153}$$

T will play the role of stretching $\Delta\eta$ time intervals into the corresponding $T\Delta\eta_1$ transformed time intervals, making time variations go slower for larger values of T. A more adequate way to implement this transformation is to consider a 1-parameter family of (FLRW) metrics, with a scale factor given by $a_T(\eta) \equiv a(\eta/T)$. Then, any metric-dependent functions $f(\eta) = f(a(\eta))$, such as $\omega_k(\eta)$, will accordingly transform as:

$$f_T(\eta) \equiv f(\eta/T) = f(\eta_1). \tag{3.154}$$

In practice, this transformation will "make the spacetime expansion go slower" as we take larger values of T, which will modify our dynamic equations (we can recover the original equations taking $T = 1$) by diminishing the relative magnitude of terms associated with time variations (time derivatives) of metric-dependent quantities. More precisely:

$$\frac{d}{d\eta}f\left(\frac{\eta}{T}\right)=\frac{1}{T}\frac{d}{d\eta_1}f(\eta_1)\equiv\frac{1}{T}f'(\eta) \qquad\Rightarrow\qquad \frac{d^n}{d\eta^n}f\left(\frac{\eta}{T}\right)=\frac{1}{T^n}f^{(n)}(\eta_1). \tag{3.155}$$

Particularly, as $T \to \infty$ all terms that contain any time derivatives of the metric vanish, producing the so called adiabatic limit. Terms with different powers of T^{-1} will decay at different rates, which we can use to hierarchize different contributions in function of slowness. Thus, we refer to terms proportional to T^{-n} as *nth* adiabatic order terms; in practice, the adiabatic order will be simply a count of time derivatives, as we can see in (3.155). With this hierarchy in mind, we can recursively compute an asymptotic series for W_k in equation (3.151), starting from the 0th adiabatic order solution $(W_k)^{(0)}$:

$$\Big((W_k)^{(0)}(\eta_1)\Big)^2=\omega_k^2(\eta_1). \tag{3.156}$$

Iterating this at (3.151), we obtain the 2nd order solution for W_k:

$$\Big(W_k^{(2)}(\eta_1)\Big)^2=\omega_k^2(\eta_1)-\frac{1}{2T^2}\left(\frac{\ddot{\omega}_k(\eta_1)}{\omega_k(\eta_1)}-\frac{3}{2}\frac{\dot{\omega}_k^2(\eta_1)}{\omega_k^2(\eta_1)}\right), \tag{3.157}$$

such that W_k will differ from $W_k^{(2)}$ only by terms of 3rd adiabatic order, or higher. In fact, as we can see from eq. (3.151), successive iterations produce only terms of even adiabatic order, all the odd-order terms vanishing identically. We can then write $W_k = W_k^{(2)} + \mathcal{O}(T^{-4})$.

Illustrating the procedure a little further, we write the calculation results to the 4th order term:

$$\begin{aligned}\Big(W_k^{(4)}\Big)^2 &= (W_k^{(2)})^2-\frac{1}{2T^2}\left(\frac{\ddot{W}_k^{(2)}}{W_k^{(2)}}-\frac{3}{2}\frac{(\dot{W}_k^{(2)})^2}{(W_k^{(2)})^2}\right)\\ &=\omega_k^2-\frac{1}{2T^2}\left(\frac{\ddot{\omega}_k}{\omega_k}-\frac{3}{2}\frac{\dot{\omega}_k^2}{\omega_k^2}\right)+\frac{1}{8T^4}\left[\frac{\ddddot{\omega}_k}{\omega_k^3}-10\frac{\dot{\omega}_k\dddot{\omega}_k}{\omega_k^4}-\frac{11}{2}\frac{\ddot{\omega}_k{}^2}{\omega_k^2}+\frac{93}{2}\frac{\dot{\omega}_k{}^2\ddot{\omega}_k}{\omega_k^5}-\frac{279}{8}\frac{\dot{\omega}_k{}^4}{\omega_k^6}\right].\end{aligned} \tag{3.158}$$

These adiabatic expansions then provide us with a natural (approximate) generalization for the concept of vacuum and particles to fully dynamic spacetimes. The case where we had asymptocally flat regions is seen to be a case where the zeroth order adiabatic approximation becomes asymptotically exact. For other, dynamical regions of spacetime, there will always be *exact* solutions of the field equations which we can *locally* identify as positive frequency by matching them with the positive-frequency adiabatic expansion at a given time. Particularly, in our separable FLRW case, for which each value of $\mathbf{k}$ will be only associated with two linearly independent solutions, $\{u_{\mathbf{k}}, (u_{\mathbf{k}})^*\}$, we can identify a positive frequency mode with an *Ath*-order adiabatic approximation $u_{\mathbf{k}}^{(A)}$ as:

$$u_{\mathbf{k}}(\mathbf{x}, \eta) = \alpha_{\mathbf{k}}^{(A)}(\eta) u_{\mathbf{k}}^{(A)}(\mathbf{x}, \eta) + \beta_{\mathbf{k}}^{(A)}(\eta)(u_{\mathbf{k}}^{(A)})^*(\mathbf{x}, \eta), \tag{3.159}$$

where:

$$\alpha_{\mathbf{k}}^{(A)}(\eta) = 1 + \mathcal{O}(T^{-(A+1)}), \tag{3.160a}$$

$$\beta_{\mathbf{k}}^{(A)}(\eta) = 0 + \mathcal{O}(T^{-(A+1)}). \tag{3.160b}$$

Here, we can match the solutions for a given time η_0 by identifying, $u_{\mathbf{k}}(\mathbf{x}, \eta_0) = u_{\mathbf{k}}^{(A)}(\mathbf{x}, \eta_0)$. However, making the identifications at different times will generally yield different exact modes $u_{\mathbf{k}}$.

This is entirely analogous to when we write the expansions of the positive frequency modes in asymptotically flat regions in terms of one another (particularly, in the 'quasidiagonal' FLRW case):

$$u_{\mathbf{k}}^{(f)} = \alpha_{\mathbf{k}} u_{\mathbf{k}}^{(p)} + \beta_{\mathbf{k}} (u_{\mathbf{k}}^{(p)})^*, \tag{3.161}$$

only in this latter case all nonzero adiabatic orders asymptotically vanish in the remote regions, making $u_{\mathbf{k}}^{(p)}$ and $u_{\mathbf{k}}^{(f)}$ asymptotic approximations up to infinite order.

In the general, dynamical regime, we can use the exact modes matched to an adiabatic approximation $u_{\mathbf{k}}^{(A)}$ to define a vacuum state $|0^{(A)}\rangle$. Although this state is highly nonunique, and inertial observers will generally measure particles in them, the particle content for *any* adiabatic vacuum states will be suppressed at high energies (at least as fast as $k^{-(A+1)}$ or $m^{-(A+1)}$ (1)). Thus, these states allow for an approximate generalization of the concept of vacuum in dynamical spacetimes, which will display a consistent behaviour in the limit of arbitrarily high frequencies. In the next chapter, we shall see that these adiabatic expansions will play a key role in the renormalization of UV divergences for our theory in curved spaces, particularly in FLRW spaces.

Although these expansions are only asymptotic (meaning they will generally diverge, rather than converge to a solution), the asymptotic expansion of a function (in our case, of an exact solution) will be unique. The converse, however, is not generally true; a particular asymptotic expansion may represent more than one function (more than one exact solutions). This will reflect in the fact that the determination of a vacuum state will not be unique in dynamical spacetimes, even if we constrain our exact modes up to arbitrarily high adiabatic orders.

4 REGULARIZATION AND RENORMALIZATION

After going through the basic procedures of quantization for free fields in a classical curved background spacetime, and exploring some of its most direct physical aspects and physical consequences, we now turn our attention to the more delicate and intricate problem of handling formally divergent quantities in our quantized theory, and making physical sense out of them.

As we can already anticipate from the energy divergences in flat spacetime – which had to be properly circumvented to calculate the Casimir Effect (see section 2.3) – certain key observables in our theory will be plagued by divergences. In fact, as we shall demonstrate briefly, these divergences are generally worse for fields quantized in a curved spacetime background than their flat counterparts, even for free (noninteracting) fields; it turns out that the implicit interactions with gravity give rise to extra divergent terms.

As we have discussed in Appendix A, the appearence of divergences is not surprising whenever we are dealing with observables quadratic in field amplitudes, such as $\phi^2(x)$ or $T_{\mu\nu}(x)$. Nonetheless, such observables are a vital portion of the dynamical elements of our theory, and if we are to make full sense of them and derive physical predictions, we must find a suitable way to modify these formally divergent expressions in order to obtain finite physical results.

In this chapter, we attack the intricate problem of renormalization as follows: first, in section 4.1, we underline some fundamental remarks in the nature of this problem, illustrating how divergences in curved space are generally worse than in flat spaces, and briefly mentioning how these may be reabsorbed in the definition of gravitational parameters in semiclassical gravity; this shall be the basis for a more general approach to renormalization in curved spaces. However, due to the more convoluted nature of this approach, we postpone its discussion to section 4.3[A]. In section 4.2 we present a rather practical and more physically intuitive renormalization scheme: adiabatic subtraction. Then, in section 4.3, we present a brief introduction to Lagrangian approaches to quantum theory, both in the form of Feynman path integrals and of the Schwinger action principle, and use them to derive the effective action. We then exhibit the divergences of the effective action, and isolate them in just a finite number of geometrical terms in an asymptotic expansion, showing that it can be be rendered finite by a renormalization of geometrical parameters in a semiclassical theory of gravity.

[A] This is by no means the most *logical* presentation sequence, but it will allow one to ‘get to the physics’ more quickly and develop some level of intuition and operational experience in this intricate subject, before dwelling into more complicated calculations.

4.1 Divergences in Curved Spaces and Semiclassical Gravitation

As we have stated many times before, the values obtained for many formal expressions quadratic in field operators are in general divergent. Even in the simplest example of the 'standard' vacuum in Minkowski spacetime, expected values such as $\langle 0|T_{\mu\nu}(x)|0\rangle$ and $\langle 0|\phi^2(x)|0\rangle$ present ultraviolet divergences.

In the more particular case of flat spacetimes, it is typically possible to renormalize the values of vacuum energy (either in nontrivial topologies or in the presence of flat boundaries[B]) by subtracting the "Minkowski vacuum corresponding value", which is taken as a reference for null energy density. In curved spacetimes, however, this procedure is generally not possible. A first reason is that, while in flat spacetimes only *energy differences* are directly observable, in General Relativity (and, to an appropriate extent, in QFTCS) *absolute energy values* appear as sources for spacetimes curvature. Furthermore, the implicit gravitational interactions may cause additional divergences; qualitatively, this is much similar to the case of free vs. interacting field theories in Minkowski spacetime where there are fundamental differences between the asymptotic limits of a weakly interacting theory[C] (*e.g.*, in the gravitational case, taking $G \to 0$) and its "free" counterpart (*e.g.*, taking $G = 0$ to start with).

Following (1), let us then show a simple example to illustrate the extra divergences that arise due to spacetime curvature. We consider a conformally flat FLRW space, whose scale factor a is defined by:

$$a(t) = \sqrt{1 - A^2t^2}, \qquad A = cte, \qquad |t| < A^{-1}, \tag{4.1}$$

and a massless scalar field ϕ, minimally coupled to gravity, whose equations of motion are simply $\Box\phi = 0$.

Using its differential definition, $d\eta = \pm a(t)dt$, it is easy to obtain conformal time η as a function of proper time t, and vice-versa; the two are related by:

$$A\eta = \arcsin(At) \quad \Leftrightarrow \quad At = \sin(A\eta) \quad \Rightarrow |\eta| < -\frac{\pi}{2A}. \tag{4.2}$$

We then have the metric components in conformal coordinates:

[B] Curved boundaries in flat space turn out to be a more complicated issue. For a further account of that matter see the last section of chapter 5 of (3), and references therein.

[C] For more complete accounts on that point, see *e.g.* (6) for a textbook introduction on interacting fields or (27) for a critical collection of founding papers in the field.

$$g^{00} = -g^{ii} = a^{-2}(\eta) = \cos^{-2}(A\eta), \tag{4.3}$$

$$|g| = a^{8}(\eta) = \cos^{8}(A\eta), \tag{4.4}$$

from which it is simple to compute the D'Alembertian: $\Box\phi = |g|^{1/2}\partial_\mu\Big(|g|^{1/2}g^{\mu\nu}\partial_\nu\phi\Big)$. Using our well-known ansatz, $u_{\mathbf{k}}(\mathbf{x},\eta) = e^{i\mathbf{k}\,\cdot\,\mathbf{x}}\chi_k(\eta)$, we obtain the equation:

$$\ddot{\chi}_k + 2H\dot{\chi}_k + k^2\chi_k = 0, \tag{4.5}$$

where H denotes the fractional expansion rate $\dot{a}/a$. In conformal time, this yields simply $H = -A\tan(A\eta)$.

Then, defining $h_k = a^{-1}\chi_k$, we arrive at a simple (*time-independent*) harmonic oscillator:

$$\ddot{h}_k + (A^2 + k^2)h_k = 0. \tag{4.6}$$

Thus, implementing a proper normalization, we arrive at the complete field modes:

$$u_{\mathbf{k}} = (16\pi^3)^{-1/2}a^{-1}(\eta)(A^2+k^2)^{-1/2}e^{i\mathbf{k}\,\cdot\,\mathbf{x}-i(A^2+k^2)^{1/2}\eta}, \tag{4.7}$$

in terms of which we can write the field expansion (3.31).

Having obtained a quantized field expansion, the main goal of our analysis is to compute the expected values of energy-momentum observables and identify their divergent terms. In this simple case of a massless, minimally coupled scalar field, the stress tensor is given simply by[D]:

$$T_{\mu\nu} = \nabla_\mu\phi\,\nabla_\nu\phi - \tfrac{1}{2}g_{\mu\nu}\nabla_\alpha\phi\nabla^\alpha\phi. \tag{4.8}$$

Further, we have that $\nabla_\mu\phi = \partial_\mu\phi$. Then, particularly, we have the energy density operator:

$$T_0^{\;0} = \tfrac{1}{2}\Big(\partial_0\phi\,\partial^0\phi - \partial_i\phi\,\partial^i\phi\Big) = \tfrac{1}{2}g^{00}\Big(\dot{\phi}^2 + (\nabla\phi)^2\Big), \tag{4.9}$$

[D] We shall analyze the stress tensor in more detail in section 4.2.3. There is a subtlety in taking the massless limit for a quantized field, but we shall ignore it at this point, as it is irrelevant to the structure of the UV divergences in which we are interested at this point.

where we have used (4.3) in the last equality.

Let then $|0\rangle$ be the vacuum state associated to the modes (4.7). We may compute its corresponding vacuum energy density:

$$\begin{aligned}\rho(x) = \langle 0|T_0{}^0(x)|0\rangle = \tfrac{1}{2}g^{00}(x)\int d^3\mathbf{k}\dot{u}_{\mathbf{k}}(x)\dot{u}^*_{\mathbf{k}}(x) + (\nabla u_{\mathbf{k}}(x))\cdot(\nabla u^*_{\mathbf{k}}(x))\\ = \frac{1}{32\pi^3 a^4(\eta)}\int d^3\mathbf{k}\Big[(k^2+A^2)^{1/2} + (k^2+H^2)(k^2+A^2)^{-1/2}\Big].\end{aligned} \tag{4.10}$$

From the asymptotic form of the integrand in (4.10) as $k \to \infty$, one can easily see that this energy density diverges quartically in the UV. Similarly to what we did in the Casimir effect, we may keep track of the divergent terms by introducing a regularizer $e^{-\alpha(k^2+A^2)^{1/2}}$. For convenience, we also multiply both sides of (4.10) by $a^4(\eta)$[E], obtaning the following result:

$$\begin{aligned}\rho a^4 &= \frac{1}{32\pi^3}\,4\pi\int_0^\infty dk\, e^{-\alpha(k^2+A^2)^{1/2}}k^2\Big[(k^2+A^2)^{1/2} + (k^2+H^2)(k^2+A^2)^{-1/2}\Big]\\ &= \frac{1}{32\pi^2}\,4\int_A^\infty d\omega\, e^{-\alpha\omega}\omega(\omega^2-A^2)^{1/2}\Big[\omega + (\omega^2-A^2+H^2)\omega^{-1}\Big]\\ &= \frac{1}{32\pi^2}\Bigg\{4\int_A^\infty d\omega\, e^{-\alpha\omega}\Big[2\omega^3 + (H^2-2A^2)\omega + \tfrac{1}{2}A^2(H^2-A^2)\omega^{-1} + \mathcal{O}(\omega^{-3})\Big]\Bigg\}.\end{aligned} \tag{4.11}$$

The regularizer has allowed us to temporarily tame the divergences in the first 3 terms, which are quartic, quadratic and logarithmic, respectively. Carrying integrations by parts for the first two terms and making a change in variables in the third one ($\omega \to \alpha\omega$), one may then arrive at the expansion:

$$\rho a^4 = \frac{e^{-\alpha A}}{32\pi^2}\Big[\frac{48}{\alpha^4} + \frac{4H^2-8A^2}{\alpha^2} + 2A^2(H^2-A^2)\ln(\alpha) + \mathcal{O}(\alpha^0)\Big]. \tag{4.12}$$

Of course, if we relax the regularization, letting $\alpha \to 0^+$, we reobtain a quartic, a quadratic and a logarithmic divergent terms. Note that, in the limit of a Minkowski spacetime $A, H \to 0$, this vacuum energy is still divergent, due to the quartic term. Nevertheless, we see explicitly that spacetime curvature has led to the emergence of *additional* quadratic and logarighimic divergences. Thus, in general, *one cannot obtain a finite energy in curved spaces just by subtracting a Minkowski-vacuum contribution.*

[E] One may think of the observable in the LHS of this equation as the (classically conserved) quantity associated to the energy in a coexpanding unitary volume $1 \times a^3$ corrected by the cosmological redshift factor $\times a$. See the next chapter (and subsection 4.2.3) for more details on this point.

An alternative strategy to handle the infinities in $\langle T_{\mu\nu} \rangle$ in curved spacetimes is presented when we consider not only a theory of quantized fields propagating in a *fixed* background geometry, but rather a wider dinamical theory which couples quantized fields to a *classical, but dynamical* spacetime: semiclassical gravitation. In this approach, one attempts to incorporate the gravitational backreaction of the quantum fields under consideration, by coupling the *expected values* of energy-momentum currents as the source of spacetime curvature.

In the purely classical case, we had Einstein's equations (3.2) coupling $T_{\mu\nu}$ to spacetime curvature:

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R + \Lambda g_{\mu\nu} = -8\pi G T_{\mu\nu}. \tag{4.13}$$

If we quantize the matter fields alone, the proposed anologue in this semiclassical scheme is to substitute $T_{\mu\nu}$ by $\langle T_{\mu\nu} \rangle$, so that we have:

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R + \Lambda_B g_{\mu\nu} = -8\pi G_B \left\langle T_{\mu\nu} \right\rangle. \tag{4.14}$$

For the time being, (4.14) is merely a formal equation, since $\langle T_{\mu\nu} \rangle$ is generally divergent. The rough idea of renormalization in this wider theoretical framework is to *absorb* its infinities by redefining the theory's so-called *bare parameters* (such as Λ_B or G_B) into new, renormalized ones, which will appear in the equations with the finite 'physical' part of $\langle T_{\mu\nu} \rangle$:

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R + \Lambda g_{\mu\nu} = -8\pi G \left\langle T_{\mu\nu} \right\rangle_{phys}. \tag{4.15}$$

In order to achieve this renormalization in a systematic way, one first turns to a more fundamental object in the theory, the action, in terms of which $T_{\mu\nu}$ in classically defined. Recall that we define the classical action S with a purely geometrical term and a matter one:

$$S = S_G + S_M,$$

with $\frac{\delta G}{\delta g^{\mu\nu}} = 0$ implying the Einstein equation above, such that the matter stress tensor is defined as:

$$\frac{2}{\sqrt{-g}} \frac{\delta S_M}{\delta g^{\mu\nu}} = T_{\mu\nu}. \tag{4.16}$$

Then, in the semiclassical theory, we seek for an object W, which we will call the *effective action*, whose functional derivative with respect to the metric yields:

$$\frac{2}{\sqrt{-g}}\frac{\delta W}{\delta g^{\mu\nu}} = \langle T_{\mu\nu}\rangle \,, \tag{4.17}$$

where a more precise definition of the braket $\langle ...\rangle$ will be given in section 4.3.

To find an object W that satifies the relation (4.17), we will present in section 4.3 the Schwinger action principle formulation to quantum mechanics, which is based on a Lagrangian formalism and is intimately related to Feynman path integrals. Then, in the context of field theory, we will find that W is generally divergent, and that one way to obtain finite quantities for observables of the quantized matter fields will be to renormalize W; this can be achieved by absorbing its divergent portion W_{div} *in the geometrical action* S_G, redefining (renormalizing) its basic geometrical parameters.

At this point, however, we shall postpone the treatment of a great such an intricate approach to renormalization and explore a simpler and subtraction scheme, working directly with the spectral representation of the stress tensor: adiabatic subtraction. In the next section, we will develop this aproach comprehensively, and use it both to obtain some operational intuition with renormalization in curved spaces, and to derive physical results of interest, such as the renormalized stress tensor in FLRW spacetimes.

4.2 Adiabatic Subtraction

The method of renormalization that we shall concretely develop in this dissertation is the one called *adiabatic subtraction*. It makes thorough use of the adiabatic expansions presented in section 3.5, which are employed both conceptually, insofar as the adiabatic condition allows us to separate positive- and negative-frequency modes and obtain a useful notion of a vacuum state, and operationally, as they are used to subtract UV-divergent terms and yield a finite result for the expectation values of various observables.

In general lines, the procedure consists of a mode-by-mode subtraction of divergent terms in the formal expression of several observables, by representing this expression in an asymptotic expansion, identifying in this expansion the terms which contain UV-divergent terms when summed (or integrated) up to arbitrarily high momenta, and then subtract such terms *inside the integration*.

In order to carry that procedure, we begin by exposing a slightly different construction for adiabatic expansions as in section 3.5, by making these expansions in proper time, as well as explicitly computing the *nth* adiabatic order solutions $h_n(t)$ before turning to WKB frequencies $W^{(n)}$. We then show how successive adiabatic orders follow a well defined hierarchy in the divergences of observables, whereupon we can design a prescription to systematically eliminate these divergences by subtracting a finite number of terms in the adiabatic expansion. Finally, we shall apply this prescription to compute the power spectrum, as well as the stress tensor for a scalar field, and interpret the obtained results.

4.2.1 Adiabatic Expansion in proper time

As in the previous approach, our starting point is to write *exact* field mode solutions (conveniently separated in this spatially flat FLRW spacetime) in the form:

$$f_{\mathbf{k}}(x) = \frac{1}{\sqrt{Va^3(t)}} e^{i\mathbf{k}\,\cdot\,\mathbf{x}} h_{\mathbf{k}}(t), \tag{4.18}$$

for which the dynamic equations $[\Box + m^2 + \xi R] f_{\mathbf{k}} = 0$ also yield a time-dependent harmonic oscillator for the temporal amplitude $h_{\mathbf{k}}$:

$$\frac{d^2 h_{\mathbf{k}}}{dt^2} + \Omega_k^2(t) h_{\mathbf{k}} = 0. \tag{4.19}$$

Only in this form, it has different expression for the frequency Ω_k, which we write in the form:

$$\Omega_k^2 = \omega_k^2 + \sigma, \tag{4.20}$$

where we define:

$$\omega_k \equiv \sqrt{\frac{k^2}{a^2} + m^2}\,, \qquad \sigma \equiv \left(6\xi - \frac{3}{4}\right)\frac{\dot{a}^2}{a^2} + \left(6\xi - \frac{3}{2}\right)\frac{\ddot{a}}{a}, \tag{4.21}$$

splitting Ω_k in an "instantaneous frequency", ω_k, and an extra contribution due to $\sqrt{\sigma}$, associated with the time variation of the scale factor of the universe $a(t)$ (note that σ has contributions both from the direct coupling with the curvature, which can be found in terms proportional to ξ, and from time derivatives of the frequency $\dot{\omega}_k$).

Of course, in the limit of a static Minkowski spacetime, $a \to cte$ (more precisely, $\dot{a}/a \to 0$, and all expressions obtained by subsequent time derivatives of it), we recover the familiar plane-wave solutions: $h_k = (2\omega_k)^{-1/2} e^{-i\omega_k t}$. But in general, for a non-static spacetime, we have the more complicated time-dependent harmonic oscillator (4.19).

Now, just as we did in section 3.5, we want to analyze (4.19) in the limit of an arbitrarily slow expansion/variation, in order to build an asymptotic series to h_k, starting from the zeroth order plane-wave approximation. Once again, we introduce the adiabatic parameter T, substituting any metric-dependent functions $f(t)$ by $f_T(t) = f(t/T)$, which can be thought of as a time rescaling:

$$t \to t' = \frac{t}{T} \qquad \Rightarrow \qquad \frac{d^n f}{dt^n}\Big(\frac{t}{T}\Big) = \frac{1}{T^n} f^{(n)}(t'), \tag{4.22}$$

(such that, in all metric dependent functions, the coordinate time shifts will be rescaled as $t \to tT$).

In practice, all these changes are mediated by the transformation in the scale factor $a(t) \to a_T(t) = a(t/T)$. We then note that one could generalize this approach to arbitrary metrics by reescaling *all* spacetime distances as $x \to xT$ (*e.g.*, rescaling geodesic distances by means of Riemann normal coordinates), or, as we do above, by transforming the metric (and any metric-dependent functions) as $g_{\mu\nu}(x) \to {}^T g_{\mu\nu}(x) = g_{\mu\nu}(x/T)$. Then, as we take $T \to \infty$ we are effectively stretching spacetime distances in all directions, and diluting any effects of curvature.

Returning to our discussion in FLRW spaces, a quite direct way to implement successive adiabatic approximations is to define iterative variable changes, starting with:

$$t_1 = \int^t dt' \Omega(t'), \tag{4.23}$$

$$h_1 = \Omega^{1/2} h. \tag{4.24}$$

Then, it is relatively straightforward to find the form of our equations (4.19) in the transformed variables. From the differential relation $dt_1 = \Omega(t) dt$, we have that:

$$\frac{df}{dt} = \Omega(t) \frac{df}{dt_1}. \tag{4.25}$$

Thus:

$$\begin{aligned} 0 &= \frac{d^2}{dt^2} h + \Omega^2 h \\ &= \Omega \frac{d}{dt_1}\Big(\Omega \frac{d}{dt_1}(\Omega^{-1/2} h_1)\Big) + \Omega^{3/2} h_1 \\ &= \Omega^{3/2}\Big\{ h_1'' + \Big(\frac{1}{4}\frac{\Omega'^2}{\Omega^2} - \frac{1}{2}\frac{\Omega''}{\Omega} + 1\Big) h_1 \Big\}, \end{aligned} \tag{4.26}$$

where the primes $'$ denote differentiation with respect to t_1. We can then rewrite the last equation as:

$$\frac{d^2}{dt_1^2}h_1 + \Omega_1^2 h_1 = 0, \tag{4.27}$$

where we have defined:

$$\Omega_1^2 \equiv 1 + \epsilon_2, \qquad \epsilon_2 = \frac{1}{4}\frac{\Omega'^2}{\Omega^2} - \frac{1}{2}\frac{\Omega''}{\Omega} = \Omega^{-1/2}\frac{d^2}{dt_1^2}\Omega^{1/2}. \tag{4.28}$$

Since each derivative of any function with respect to t_1 is proportional to its derivative with respect to t (or, more loosely speaking, since an infinitesimal variation in t_1 is proportional to an infinitesimal variation in t), we can immediately assert that ϵ_2 contains only terms of 2nd adiabatic order or higher. In the lowest (zeroth) order, equation (4.27) is merely a time-independent oscillator, whose linearly independent solutions are:

$$h_1(t) \propto e^{\mp it_1} + \mathcal{O}(T^{-2}) = e^{\mp i\int^t \Omega(t')dt'} + \mathcal{O}(T^{-2}). \tag{4.29}$$

Evaluating the solution in the limit of an infinitely slow expansion ($T \to \infty$), we take the positive-frequency solutions in a general FLRW spacetime to be those which match the usual Minkowski positive-frequency solutions, namely, the ones with a minus sign on the exponent.

Upon comparison of equations (4.19) and (4.27), we see that they are formally identical in their respective parameters. Such self-similarity allows us to easily define iterations to obtain higher adiabatic orders:

$$t_2 = \int_1^t dt_1'\Omega_1(t_1'), \tag{4.30}$$

$$h_2 = \Omega_1^{1/2}h_1, \tag{4.31}$$

whence it follows immediately that:

$$\frac{d^2}{dt_2^2}h_2 + \Omega_2^2 h_2 = 0, \tag{4.32}$$

being:

$$\Omega_2^2 = 1 + \epsilon_4, \quad \epsilon_4 \equiv \Omega_1^{-1/2}\frac{d^2}{dt_2^2}\Omega_1^{1/2}. \tag{4.33}$$

Since the derivatives act only on the higher order terms (cancelling the constant, zeroth order one), we have that ϵ_4 only contains terms of 4th adiabatic order or higher:

$$\begin{aligned}\epsilon_4 &= (1+\epsilon_2)^{-1/2}\frac{d^2}{dt_2^2}(1+\epsilon_2)^{1/2} \\ &= \Big(1-\frac{1}{2}\epsilon_2+\mathcal{O}(T^{-4})\Big)\frac{d^2}{dt_2^2}\Big(\overset{0}{\cancel{1}}+\frac{1}{2}\epsilon_2+\mathcal{O}(T^{-4})\Big) \\ &= \frac{1}{2}\frac{d^2}{dt_2^2}\epsilon_2+\mathcal{O}(T^{-6}). \end{aligned} \tag{4.34}$$

Again, we can immediately write the solution up to 4th adiabatic order:

$$h_2 \propto e^{\mp it_2}+\mathcal{O}(T^{-6}), \tag{4.35}$$

and this substitution process can be repeated up to any desired adiabatic order, recursively defining t_n, h_n, Ω_n, which will obey analogous equations to (4.19):

$$t_n = \int_{n-1}^{t} dt'_{n-1}\Omega_{n-1}(t'_{n-1}) \tag{4.36}$$

$$h_n = \Omega_{n-1}^{1/2}h_{n-1} \tag{4.37}$$

$$\frac{d^2}{dt_n^2}h_n+\Omega_n^2h_n = 0 \tag{4.38}$$

$$\Omega_n = 1+\epsilon_{2n} \tag{4.39}$$

$$\Rightarrow \qquad h_n \propto e^{-it_n}+\mathcal{O}\Big(T^{-2(n+1)}\Big). \tag{4.40}$$

To make use of this expansion to perform *adiabatic subtractions*, it will be particularly useful to compute the expansions of the WKB frequency W_k, as well as some elementary functions of it. The iterations for W here are entirely analogous to those in section 3.5; the only difference is that, because we are working with proper time t, the frequency that appears in eq. (4.19) is not $\omega(t)$ but rather $\Omega(t)$, *which already carries contributions of second adiabatic order from* $\sigma(t)$, as σ has terms proportional to $\dot{a}^2$ and $\ddot{a}$ (see eqs. (4.20) and (4.21)).

Therefore, just as in equation (3.151), if we write $h_k = (2W_k)^{-\frac{1}{2}}\exp\Big(-i\int^t dt' W_k(t')\Big)$, eq. (4.19) yields:

$$W_k^2 = \Omega_k^2+\omega_k^{\frac{1}{2}}\frac{d^2}{dt^2}\omega_k^{-\frac{1}{2}}. \tag{4.41}$$

Before we proceed further in computing its adiabatic expansion, let us lay an unambiguous notation for W_k (in what follows, we shall often suppress the subscript k, for cleaness). As in section 3.5, we denote its truncation to *nth* adiabatic order as $W^{(n)}$. To keep track of the adiabatic order of each term, it is useful to write the expansion in the form:

$$W_k \sim \omega_k + \omega_k^{(2)} + \omega_k^{(4)} + \dots\,, \tag{4.42}$$

where we have used the fact that $\omega_k^{(0)} = \omega_k$. Thus, we write the first few truncations as:

$$\begin{aligned}
&W^{(2)} = \omega + \omega^{(2)},\\
&W^{(4)} = \omega + \omega^{(2)} + \omega^{(4)} = W^{(2)} + \omega^{(4)},\\
&\textit{etc.}
\end{aligned}$$

However, for any other functions of W, we shall use $\Big(f(W)\Big)^{(n)}$ to denote the terms of *exact adiabatic order n* in their expansion, rather than the truncation up to nth order. Thus, for example, $(W^{(2)})^2 \neq (W^2)^{(2)}$, since:

$$\begin{aligned}
(W^2)^{(2)} &\equiv \Big((\omega + \omega^{(2)} + \omega^{(4)} + ...)^2\Big)^{(2)}\\
&= \Big(\omega^2 + 2\omega\omega^{(2)} + (\omega^{(2)})^2 + 2\omega\omega^{(4)} + ...\Big)^{(2)}\\
&\equiv 2\omega\omega^{(2)},
\end{aligned}$$

which is purely of second adiabatic order, while:

$$(W^{(2)})^2 \equiv (\omega + \omega^{(2)})^2 = \omega^2 + 2\omega\omega^{(2)} + (\omega^{(2)})^2,$$

which has terms of zeroth, second and fourth adiabatic orders.

A particular class of functions $f(W)$ that we will be operating in the next section are simple powers of W, W^α, for which we have:

$$\begin{aligned}
W^\alpha &\sim \Big[\omega + \omega^{(2)} + \omega^{(4)} + ...\Big]^\alpha\\
&= \omega^\alpha\left[1 + \frac{\omega^{(2)}}{\omega} + \frac{\omega^{(4)}}{\omega} + ...\right]^\alpha\\
&= \omega^\alpha\left[1 + \alpha\Big(\frac{\omega^{(2)}}{\omega} + \frac{\omega^{(4)}}{\omega} + ...\Big) + \frac{\alpha(\alpha-1)}{2}\Big(\frac{\omega^{(2)}}{\omega} + \frac{\omega^{(4)}}{\omega} + ...\Big)^2 + ...\right].
\end{aligned} \tag{4.43}$$

From this, it is easy to group the terms of the same adiabatic order. For the 3 lowest orders, we obtain:

$$
\begin{aligned}
(W^{\alpha})^{(0)} &= \omega^{\alpha}, && \text{(4.44a)}\\
(W^{\alpha})^{(2)} &= \alpha\frac{\omega^{(2)}}{\omega}\,\omega^{\alpha}, && \text{(4.44b)}\\
(W^{\alpha})^{(4)} &= \Big[\alpha\frac{\omega^{(4)}}{\omega} + \tfrac{1}{2}\alpha(\alpha-1)\Big(\frac{\omega^{(2)}}{\omega}\Big)^2\Big]\omega^{\alpha}. && \text{(4.44c)}
\end{aligned}
$$

A case of particular interest ahead will be $\alpha = -1$, which results in:

$$
(W^{-1})^{(0)} = \omega^{-1}, \quad (W^{-1})^{(2)} = \frac{-\omega^{(2)}}{\omega^2}, \quad (W^{-1})^{(4)} = \frac{-\omega^{(4)}}{\omega^2} + \frac{(\omega^{(2)})^2}{\omega^3}. \tag{4.45}
$$

Now that we have shown how to obtain the expansions of functions of W in terms of the basic building blocks $\omega^{(n)}$, let us explicitly compute the first few orders of the latter. The zeroth order term is simply:

$$
\omega^{(0)} = \omega = \sqrt{m^2 + \frac{\mathbf{k}^2}{a^2}}. \tag{4.46}
$$

Then, recursively, it is not hard to compute the 2nd term from eq (4.41):

$$
\begin{aligned}
(W^{(2)})^2 &= \omega^2 + \sigma + \omega^{\frac{1}{2}}\frac{d}{dt^2}\omega^{-\frac{1}{2}}\\
\Rightarrow W^{(2)} &= \omega + \frac{1}{2\omega}\Big(\sigma + \omega^{\frac{1}{2}}\frac{d}{dt^2}\omega^{-\frac{1}{2}}\Big),
\end{aligned} \tag{4.47}
$$

where we have computed the square root discarding 4th or higher order terms. Thus:

$$
\begin{aligned}
\omega^{(2)} &= \frac{1}{2\omega}\Big(\sigma + \omega^{\frac{1}{2}}\frac{d}{dt^2}\omega^{-\frac{1}{2}}\Big)\\
&= \frac{1}{2\omega}\bigg\{\Big[6\xi - \frac{3}{4} - \frac{3}{2}\frac{k^2}{\omega^2 a^2} + \frac{5}{4}\Big(\frac{k^2}{\omega^2 a^2}\Big)^2\Big]\Big(\frac{\dot{a}}{a}\Big)^2 + \Big[6\xi - \frac{3}{2} + \frac{1}{2}\frac{k^2}{\omega^2 a^2}\Big]\frac{\ddot{a}}{a}\bigg\}
\end{aligned} \tag{4.48}
$$

The 4th order term, although equally straightforward, is quite laborious to compute manually, and prohibitively large to write down here. Nevertheless, its definition in the recursive expression

$$
(W^{(4)})^2 = \Omega^2 + (W^{(2)})^{\frac{1}{2}}\frac{d^2}{dt^2}(W^{(2)})^{-\frac{1}{2}} \tag{4.49}
$$

allows one to easily implement these calculations symbolically in a computer. We will use such results (presently computed in the software Mathematica) in the following sections to evaluate the renormalization of the stress tensor.

Now that we have extensively developed adiabatic expansions, we shall occupy ourselves in the next section in making use of them to systematically subtract infinities, and obtain finite expectation values for observables of physical interest.

4.2.2 Structure of the divergences and Adiabatic Subtraction

We already know that there will be divergences in the expectation values of many relevant physical observables; particularly, in FLRW spaces, these expectation values can be put in the form of a Fourier expansion in terms of the **k**, which takes the form of integrals of (almost everywhere) finite functions of k (i.e. finite integrands). The divergences then occur when we integrate these funtions in the UV region, $k \to \infty$ [F]. By means of the adiabatic expansion, these integrands may be split in a sum of integrable (convergent) and non-integrable (divergent) terms, and, furthermore, *this expansion has a well defined hierarchy for the order of the divergences in its terms.* (The meaning of this last sentence should become clearer below.)

The simplest example at hand is the two-point field amplitude $\phi(x)\phi(x')$. If we keep x and x' independent, this is a well defined (operator-valued) distribution (see Appendix A). If we attempt to evaluate its vacuum expectation value, we get the following expansion:

$$\langle 0|\phi(x)\phi(x')|0\rangle = \sum_{\mathbf{k}} f_{\mathbf{k}}(x) f^*_{\mathbf{k}}(x') \to \frac{1}{2(2\pi)^3}\Big[a(t)a(t')\Big]^{-3/2} \int d^3\mathbf{k} e^{i\mathbf{k}\,\cdot\,(\mathbf{x}-\mathbf{x}')} h_{\mathbf{k}}(t) h^*_{\mathbf{k}}(t'). \tag{4.50}$$

(The vacuum state $|0\rangle$ considered here is that determined by the modes $f_{\mathbf{k}}$. Among other things, it will be approximated by any adiabatic vacuum $|0^{(A)}\rangle$.)

Although this does not properly yield a function of (x, x'), since (4.50) does not absolutely converge, it is well defined as a (number-valued) distribution, and can be straightforwardly evaluated inside integrals (as we did in last chapter for obtaining the response function of particle detectors). However, we are in much more serious trouble when trying to evaluate the integral (4.50) by itself as a function of spacetime, making $x = x'$. Formally, we write the expansion:

$$\langle 0|\phi^2(x)|0\rangle = \frac{1}{2(2\pi)^3 a^3(t)} \int d^3\mathbf{k} |h_k(t)|^2 = \frac{1}{4\pi^2 a^3(t)} \int_0^\infty dk k^2 |h_k(t)|^2. \tag{4.51}$$

If we check the ultraviolet limit of this integrand $k \gg m, \sigma$, we obtain:

[F] Sometimes, there may also be divergences in the IR, $k \to 0$, but these are generally not so pervasive; they usually appear for specific ranges of parameter in the theory and can be highly dependent on the choice of vacuum state. We shall not systematically occupy ourselves with them in the scope of this text.

$$k^2|h_k|^2 \sim k^2\omega_k^{-1} \sim k, \tag{4.52}$$

(where we are ignoring the k-independent scale factor $a(t)$).

Therefore, if we set a very large UV cutoff K for this integral, we will have:

$$\int_0^K dkk^2|h_k(t)|^2 \sim \mathcal{O}(K^2). \tag{4.53}$$

This means that this observable diverges quadratically. To be more precise, its dominant term diverges quadratically. Had we considered a higher order expansion of ω_k^{-1}, we would obtain:

$$\begin{aligned}\omega_k^{-1} &= \left[\frac{k^2}{a^2} + m^2\right]^{-\frac{1}{2}} \\ &= \frac{a}{k} - \frac{1}{2}\frac{m^2a^3}{k^3} + \frac{3}{8}\frac{m^4a^5}{k^5} + \dots \ .\end{aligned} \tag{4.54}$$

In that case, the integral above yields:

$$\int_0^K dkk^2\omega_k^{-1} \sim \mathcal{O}(K^2) + \mathcal{O}(\ln K) + \mathcal{O}(K^{-2}) + \dots \ , \tag{4.55}$$

where we can see that it has quadratic *and* logarithmic divergences, as well as a series of convergent terms. If we now analyze the first few terms in the adiabatic expansion for $|h_k(t)|^2 = W_k^{-1}$, we can verify that *the dominant UV contribution decreases in direct corresponce with the adiabatic order*, i.e.:

$$(W_k^{-1})^{(0)} \sim k^{-1}, \tag{4.56a}$$

$$(W_k^{-1})^{(2)} \sim k^{-3}, \tag{4.56b}$$

$$(W_k^{-1})^{(4)} \sim k^{-5}, \tag{4.56c}$$

and, generally, $(W_k^{-1})^{(2n)} \sim k^{1-2n}$ (although for some very special parameter values (ξ, m), the coefficients for the highest power terms may turn out to be 0). Thus, if we evaluate the adiabatic expansion for the $\langle\phi^2\rangle$ integral (4.51), we obtain:

$$\begin{aligned}\int_0^K dkk^2|h_k(t)|^2 &= \int_0^K dkk^2W_k^{-1} \\ &\sim \int_0^K dkk^2\Big((W_k^{-1})^{(0)} + (W_k^{-1})^{(2)} + (W_k^{-1})^{(4)} + \dots\Big) \\ &= \mathcal{O}(K^2) + \mathcal{O}(\ln K) + \text{convergent terms,}\end{aligned} \tag{4.57}$$

where the divergences have spawned only from the zeroth and the second order adiabatic terms: $(W_k^{-1})^{(0)}$ and $(W_k^{-1})^{(2)}$. More especifically, $(W_k^{-1})^{(0)}$ yields *both* quadratic and logarithmic divergences (+ convergent terms), while $(W_k^{-1})^{(2)}$ yields *only* logarithmic divergences (+ convergent terms).

Thus, one way to get rid of all infinities and obtain a finite expectation value for $\langle\phi^2\rangle$ – to which we shall ascribe physical meaning and compare with experiments – is to *subtract from* $|h_k|^2 = W_k^{-1}$ *all terms up to second adiabatic order inside the integration sign*, and then carry the integration for a convergent integrand. That is, we *define*:

$$\langle 0|\phi^2(x)|0\rangle_{phys} \equiv \frac{1}{4\pi^2 a^3(t)} \int_0^\infty dk k^2 \Big[|h_k(t)|^2 - (W_k^{-1})^{(0)}(t) - (W_k^{-1})^{(2)}(t)\Big]. \tag{4.58}$$

Another very important bilinear observable in field theories is the stress tensor $T_{\mu\nu}$, which, as we have said before, conveys information about the energy and momentum of the field. Besides its bilinear form in field operators, some of its terms also contain second spacetime derives, or are quadratic in first derivatives. These derivatives give rise to two extra powers of k (or ω_k, which behaves as $\sim k$ in the UV), making the divergences on $T_{\mu\nu}$ worse than those in ϕ^2: besides logarithmic and quadratic divergent terms, it also contains quartic divergences (indeed, we have already come across them in the previous section; see eq (4.12)). Therefore, in order to eliminate all divergences in $\langle T_{\mu\nu}\rangle$, one must generally make subtractions up to 4th adiabatic order.

If we are to generalize this procedure to any observable Q in our theory, and we wish to make it sufficiently systematic to obtain a unique physical prediction (less of residual free renormalization parameters, whose values should be experimentally determined), there are a couple of things we should pay attention to. First, different observables may, quite naturally, present different types of divergency and therefore require different orders of adiabatic subtraction to be rendered finite. Second, within one single adiabatic order, one will generally find both divergent and convergent terms (as is well illustrated at eq (4.54)); in principle, one only has to subtract the divergent contributions to get a finite result, but the questions of how to decompose each adiabatic term and whether to subtract its convergent parts leave ambiguities of which finite result we will end up with. Moreover, the leading UV behaviour of each adiabatic term may depend on the parameters of the theory (in the present example of a free scalar field, nonminimally coupled to gravity, these are basically the mass m, and the adimensional coupling constant ξ); for some special values of parameters, the leading UV coefficients may turn out to be 0, making a otherwise divergent term convergent (e.g., for $\xi = 1/6$, $(W_k^{-1})^{(2)}$ decays as k^{-5}, rather than k^{-3}).

Therefore, in order to systematically eliminate divergences and obtain a well-defined finite expectation value, we *define* the procedure of adiabatic subtraction as follows (2): given an observable Q that, *for general values of parameters in the theory*, has a formal expansion for its expectation value $\langle\Psi|Q|\Psi\rangle$ with divergences up to the adiabatic order $Q^{(n)}$, then its physical expectation value $\langle\Psi|Q|\Psi\rangle_{phys}$ is defined by subtracting *in the expansion (i.e. under the integration sign) all terms $Q^{(A)}$ of adiabatic order $A \leq n$*, regardless of whether these terms have convergent contributions or whether they are divergent at all for the specific values of parameters under consideration.

For example, if we consider the power spectrum, the last divergent term is generally $(W^{-1})^{(2)}$. We can rewrite it, arranging all terms proportionally to positive powers of m and $\xi' = \xi - 1/6$:

$$\begin{aligned}(W^{-1})^{(2)} &= -\frac{\omega^{(2)}}{\omega^2}\\ &= -\frac{1}{2\omega^3}\left[\left(6\xi' - \frac{m^2}{\omega^2} + \frac{5}{4}\frac{m^4}{\omega^4}\right)\left(\frac{\dot{a}}{a}\right)^2 + \left(6\xi' - \frac{1}{2}\frac{m^2}{\omega^2}\right)\frac{\ddot{a}}{a}\right]\\ &= -\frac{3\xi'}{\omega^3}\frac{\dot{a}^2}{a^2} - -\frac{3\xi'}{\omega^3}\frac{\ddot{a}}{a} + \frac{m^2}{2\omega^5}\frac{\dot{a}^2}{a^2} + \frac{m^2}{4\omega^5}\frac{\ddot{a}}{a} - \frac{5m^4}{8\omega^7}\frac{\dot{a}^2}{a^2}. \end{aligned} \tag{4.59}$$

In this form, we immediately see that in the conformally special case $\xi = 1/6$ ($\xi' = 0$) this would have no divergent terms for the power spectrum. Still, according to our prescription, we should subtract the renormalized expression by subtracting this term as well. In fact, this will be necessary if we want out theory to depend continuously on its parameters.

This prescription also leads to some intriguing consequenses regarding how quantum field observables could differ from our classical expectations. For example, a quantity that is positive-definite such as $\phi^2(x)$ can, due to the subtractions, present negative expectation values. The same is true for the energy density $\mathcal{H}(x)$ – which we had already seen in renormalization in flat space, for the Casimir Effect. Moreover, as different observables may require different orders of fundamental subtraction, they could in principle *have different physical expectation values even when their classical expressions coincide for some particular value of parameters*. Indeed, there are known examples of this renormalization discrepancy; among them, we shall explore the so-called *trace anomaly* (or *conformal anomaly*) for the stress tensor in the next section.

In the next subsection, we shall explicitly present the application of this method to compute the stress tensor of a scalar field.

4.2.3 Vacuum Energy in Curved Space: adiabatic renormalization of the Stress Tensor

Two manifest advantages of the adiabatic subtraction procedure are that (i) it is, in a sense, more physically intuitive than other procedures in its execution, since one operates subtractions directly for the observables of interest in terms of field modes (and has a quite extensive interpretation framework for spectral amplitudes in physics), and (ii) it is extremely straightforward to compute predictions, either analytically or with the aid of symbolic/numerical tools, based on its iterative structure.

Even so, for a number of observables of interest, the necessary computations may be analytically impractical and the results, prohibitively large to even display, obscuring their physical meaning and interpretation. Particularly, this is true for the expectation values of the stress tensor, which is an essential observable for the dynamical predictions of a theory, and crucially so if we wish to explore its gravitational (and cosmological) effects. Thus, in order to properly grasp the results of adiabatic renormalization for the stress tensor, and interpret qualitative and quantitative features of vacuum energy in curved spacetimes, we start by restricting our attention to the conformally special case $m = 0$, $\xi = 1/6$; this will greatly simplify our computations, and allow us to explore fully analytical calculations. After those results have been calculated and discussed, we will explore a more general range of parameters with the aid of symbolical and numerical calculations in the next section.

We begin by computing the classical expression for the stress tensor of the scalar field (3.23), which is found by extremizing its action (3.24) with respect to the metric:

$$\begin{aligned}
T_{\mu\nu} &= \frac{2}{\sqrt{-g}}\frac{\delta S}{\delta g^{\mu\nu}} \\
&= \frac{1}{2}\frac{\partial\sqrt{-g}}{\partial g^{\mu\nu}}\Big((\nabla^\alpha\phi)(\nabla_\alpha\phi) - (m^2+\xi R)\phi^2\Big) + \frac{\sqrt{-g}}{2}\Big[\nabla_\mu\phi\nabla_\nu\phi - \xi\phi^2\Big(R_{\mu\nu} + g^{\alpha\beta}\frac{\partial R_{\alpha\beta}}{\partial g^{\mu\nu}}\Big)\Big] \\
&= \nabla_\mu\phi\nabla_\nu\phi - \tfrac{1}{2}\Big[\nabla^\alpha\phi\nabla_\alpha\phi - m^2\phi^2\Big] - \xi\Big[(R_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R)\phi^2 + \nabla_\mu\nabla_\nu\phi^2 - g_{\mu\nu}\Box\phi^2\Big], \qquad (4.60)
\end{aligned}$$

where we have used equations (3.6) and (3.8). We also stress that there will be no variations associated to first derivative terms, as ∇_μ has an invariant action on scalars (*i.e.*, $\nabla_\mu\phi = \partial_\mu\phi$).

Equation (4.60) then yields the trace:

$$\begin{aligned}
T_\mu{}^\mu &= -(\nabla_\mu\phi)(\nabla^\mu\phi) + 2m^2\phi^2 + \xi R\phi^2 + 3\xi\Box\phi^2 \\
&= (6\xi - 1)\nabla_\mu\phi\nabla^\mu\phi + m^2\phi^2, \qquad (4.61)
\end{aligned}$$

where we have used the dynamic equations for ϕ in the last line.

In this expression, if we make $m = 0$, $\xi = 1/6$, we recover the classical result that the trace of a conformally trivial field is null[G]: $T_\mu{}^\mu = 0$. For the quantized field, however, this equation must be treated with greater care. Generally, it can be problematic to impose

[G] See, *e.g.*, appendix D of (5).

classical field equalities, especially constraints, directly in terms of *operator identities*[H] (at the very least, there will be no *a priori* guarantee that they coincide with the limits $m \to 0$, $\xi \to 1/6$ of the nontrivial theory). Instead, we expect them to be implemented for the quantized theory at the level of expected values:

$$\langle \Psi \Big| \, T_{\mu}{}^{\mu}(x) \Big| \Psi \rangle \Big|_{\xi = \frac{1}{6}, m=0} = 0. \tag{4.62}$$

Indeed, equation (4.62) will hold for the formal, nonrenormalized expression of the trace. However, according to our prescription for adiabatic subtraction, we should determine the physical value of $\langle T_{\mu}{}^{\mu} \rangle$ by computing the expected value *with the appropriate (4th order) adiabatic subtraction, regardless of whether the formal expectation value is nondivergent for these specific parameters.* In fact, as $T_{\mu\nu}$ is a tensorial observable, one should actually consider $g^{\mu\nu}\langle T_{\mu\nu} \rangle$ for the expected value of the trace, and the nonrenormalized expression of $\langle T_{\mu\nu} \rangle$ carries divergences even in this conformally special case.

With some computational effort, one may verify that, indeed, even when the appropriate adiabatic subtractions are performed, the contribution to the expected value $\langle T_{\mu}{}^{\mu} \rangle$ from the term with derivatives on (4.61) vanishes when one makes $\xi \to /6$. *Thus, we take the informal liberty of writing*

$$T_{\mu}{}^{\mu} \Big|_{\xi = 1/6} = m^2 \phi^2 \tag{4.63}$$

directly at an operator level.

From (4.63), we obtain the formal expression for the vacuum expectation value:

$$\langle 0 | T_{\mu}{}^{\mu}(x) | 0 \rangle \Big|_{\xi = 1/6} = m^2 \, \langle 0 | \phi^2(x) | 0 \rangle \, . \tag{4.64}$$

However, as anticipated in last section, this does not mean that one can automatically identify the physical, renormalized expectation values:

$$\langle 0 | T_{\mu}{}^{\mu}(x) | 0 \rangle_{phys} \not\equiv m^2 \, \langle 0 | \phi^2(x) | 0 \rangle_{phys} \tag{4.65}$$

(where we have omitted the specification $\xi = 1/6$).

[H] For an illustrative example of this assertion in the context of gauge theories, see for instance chapter 5 of (6) and the implementation of the Lorenz gauge condition in the Gupta-Bleuler quantization method *as a constraint in the Hilbert space of the theory, rather than an operator identity.*

This is because ϕ^2 and $T_{\mu\nu}$ generally have different types of divergences. While the former has at most quadratic divergences, and must be subtracted only up to second adiabatic order, the latter also has quartic divergences, so that we must also subtract the fourth adiabatic term. This amounts to:

$$\langle 0|T_\mu{}^\mu(x)|0\rangle_{phys} = m^2 \langle 0|\phi^2(x)|0\rangle_{phys} - \frac{m^2}{4\pi^2 a^3(t)}\int_0^\infty dk k^2 (W_k^{-1})^{(4)}(t), \tag{4.66}$$

where the last term is the so-called *trace anomaly*; it emerges solely from the process of renormalization and will generally make the trace of a quantized theory differ from its classical counterpart. The situation is particularly interesting if we evaluate this result in the limit $m \to 0$, for which the trace classically vanishes. At first sight, both terms in (4.66) seem to vanish in this limit, as they have a m^2 prefactor; however, one must look more carefully in the spectral integrals, to see if $\langle 0|\phi^2|0\rangle_{phys}$ and $\int dkk^2(W_k^{-1})^{(4)}$ do not entail any infrared (IR) divergences that may compensate this factor.

As of the first term, it is easy to verify that both the integrals of $(W_k^{-1})^{(0)}$ and $(W_k^{-1})^{(2)}$ do not yield any IR contributions in that limit. We show that explicitly, beginning with $(W_k^{-1})^{(0)}$:

$$\begin{aligned}\lim_{m\to 0}\left\{m^2\int_0^K dk\, k^2\omega_k^{-1}\right\} &= \lim_{m\to 0}\left\{m^2\int_0^K dk\, k^2 m^{-1}\left[1+\tfrac{k^2}{m^2}\right]^{-\frac{1}{2}}\right\}\\ &= \lim_{m\to 0}\left\{m^4\int_0^K dx\, x^2[1+x^2]^{-\frac{1}{2}}\right\},\\ &= 0,\end{aligned} \tag{4.67}$$

where we have made the substitution $k \to x \equiv k/m$ to take the m dependence outside of the integral; we have also inserted a UV-cutoff K to tame the UV divergences and focus in the IR behaviour. It is then easy to see that the x-integral remains finite in the IR limit, as $x \to 0$, so that the whole expression vanishes in the massless limit.

The situation is quite similar for $(W_k^{-1})^{(2)}$. One can see from eq. (4.59) that m and ω always appear in terms with a fixed power proportion m^i/ω^{i+3}, such that the relevant integral takes the form:

$$\begin{aligned}\lim_{m\to 0}\left\{m^2\int_0^K dk\, k^2(W_k^{-1})^{(2)}\right\} &= \lim_{m\to 0}\left\{m^2\int_0^K dk\, k^2\sum_i \alpha_i^{(2)}\frac{m^i}{\omega^{i+3}}\right\}\\ &= \lim_{m\to 0}\left\{m^2\int_0^K dk\, k^2 m^{-3}\sum_i \alpha_i^{(2)}\left[1+\tfrac{k^2}{m^2}\right]^{-\frac{i+3}{2}}\right\}\\ &= \lim_{m\to 0}\left\{m^2\int_0^K dx\, x^2\sum_i \alpha_i^{(2)}\left[1+x^2\right]^{-\frac{i+3}{2}}\right\}\\ &= 0,\end{aligned} \tag{4.68}$$

where the α_i are coefficients independent of k and m. This similarly vanishes in the massless limit, although as m^2 rather than as m^4.

Then, it is not hard to anticipate what will happen with $(W_k^{-1})^{(4)}$. Although it is laborious to compute it explicitly, it is not hard to see that all its terms will follow the same tendency, bearing the fixed power proportions m^i/ω^{i+5}. Now, since this term is not UV divergent, we drop the upper cutoff $K \to \infty$ and directly show the form of its *finite* contribution to the trace:

$$\begin{aligned}
\langle 0|T_\mu{}^\mu|0\rangle_{phys} &= \frac{1}{4\pi^2 a^3}\lim_{m\to 0}\left\{m^2\int_0^\infty dk\, k^2 (W_k^{-1})^{(4)}\right\} \\
&= \frac{1}{4\pi^2 a^3}\lim_{m\to 0}\left\{m^2\int_0^\infty dk\, k^2 \sum_i \alpha_i^{(4)}\frac{m^i}{\omega^{i+5}}\right\} \\
&= \frac{1}{4\pi^2 a^3}\int_0^\infty dx\, x^2 \sum_i \alpha_i^{(4)}(1+x^2)^{-\frac{i+5}{2}} \\
&= \frac{1}{4\pi^2 a^3}\sum_i \alpha_i^{(4)}\frac{\sqrt{\pi}\,\Gamma(1+\frac{i}{2})}{4\,\Gamma(2+\frac{i}{2})},
\end{aligned} \tag{4.69}$$

as all powers of m have cancelled out in the expression in the curly brackets, yielding simply a m-independent integral.

Making use of a symbolically computed expression for $(W^{-1})^{(4)}$, we may find the explicit coefficients $\alpha_i^{(4)}$ in terms of $a(t)$ in this conformally coupled case ($\xi = 1/6$):

$$\begin{cases}
\alpha_2^{(4)} = -\dfrac{\dot{a}^4}{2a^4} - \dfrac{7\ddot{a}^2}{16a^2} - \dfrac{\ddddot{a}}{16a} - \dfrac{33\dot{a}^2\ddot{a}}{16a^3} - \dfrac{11\dot{a}\,\dddot{a}}{16a^2}, & \text{(4.70a)} \\
\alpha_4^{(4)} = \dfrac{49\dot{a}^4}{8a^4} + \dfrac{21\ddot{a}^2}{32a^2} + \dfrac{35\dot{a}^2\ddot{a}}{4a^3} + \dfrac{7a\,\dddot{a}}{8a^2}, & \text{(4.70b)} \\
\alpha_6^{(4)} = -\dfrac{231\dot{a}^4}{16a^4} - \dfrac{231\dot{a}^2\ddot{a}}{32a^3}, & \text{(4.70c)} \\
\alpha_8^{(4)} = \dfrac{1155\dot{a}^4}{128a^4}, & \text{(4.70d)}
\end{cases}$$

and obtain the following expression for the anomalous trace:

$$\langle 0|T_\mu{}^\mu|0\rangle_{phys} = \frac{1}{480\pi^2}\left[\frac{\ddot{a}^2}{a^2} + \frac{\ddddot{a}}{a} - 3\frac{\dot{a}^2\ddot{a}}{a^3} + 3\frac{\dot{a}\,\dddot{a}}{a^2}\right]. \tag{4.71}$$

For completeness, we mention that this can be computed in a generally covariant form, in terms of curvature scalars. This yields (see eq. (6.144) of (1)):

$$\langle 0|T_\mu{}^\mu|0\rangle_{phys} = -\frac{1}{2880\pi^2}\Big[R_{\mu\nu\alpha\beta}R^{\mu\nu\alpha\beta} - R_{\alpha\beta}R^{\alpha\beta} - \Box R\Big]. \tag{4.72}$$

Once again, this entails a relatively compact result after a lengthy regularization and subtraction procedure. With (4.71), one can immediately compute the trace expectation value *just from the knowledge of the spacetime expansion* $a(t)$. Note that this purely anomalous trace will actually be state-independent: had we considered a many-particle state $|\Psi\rangle = |n_{\mathbf{k_1}}, n_{\mathbf{k_2}}, \ldots\rangle$ with a finite energy difference with respect to the vacuum, we would have just an extra convergent term in $\langle\phi^2(x)\rangle$:

$$\langle 0|\phi^2(x)|0\rangle \longrightarrow \langle\Psi|\phi^2(x)|\Psi\rangle = \langle 0|\phi^2(x)|0\rangle + \sum_{\mathbf{k}} 2N_{\mathbf{k}}|u_{\mathbf{k}}(x)|^2, \tag{4.73}$$

which would yield a vanishing contribution to $\langle T_\mu{}^\mu\rangle$ as $m \to 0$.

In what follows, we will show how one is able to compute the vacuum expectation value for the entire stress tensor $\langle 0|T^{\mu\nu}|0\rangle_{phys}$ in a FLRW spacetime *just from the knowledge of its trace* $\langle T^\mu_\mu\rangle_{phys}$. In fact, as this trace is state-independent, the procedure outlined here will actually allow us to compute this expectation value *in any state* $|\Psi\rangle$ that obeys the FLRW symmetries, namely, spatial homogeneity and isotropy.

As a starting point, we note that this renormalized expectation value must obey the covariant conservation law:

$$\nabla_\mu \langle 0|T^{\mu\nu}|0\rangle_{phys} = 0. \tag{4.74}$$

This is true because (i) this equality holds for the formal, unrenormalized expectation value:

$$\nabla_\mu \langle 0|T^{\mu\nu}|0\rangle = 0, \tag{4.75}$$

and (ii) it must hold order by order in an adiabatic expansion:

$$\langle 0|T^{\mu\nu}|0\rangle \sim \langle 0|T^{\mu\nu}|0\rangle^{(0)} + \frac{1}{T^2}\langle 0|T^{\mu\nu}|0\rangle^{(2)} + \frac{1}{T^4}\langle 0|T^{\mu\nu}|0\rangle^{(4)} + \ldots \tag{4.76}$$

so that (4.75) can be true for *any* of the spacetimes in the 1-parameter family of FLRW metrics $a_T(t) = a(t/T)$ (more concisely, so that (4.75) can hold for *any* value of T). In fact, this condition is a strong motivation to define adiabatic subtractions in the way we did, subtracting *all* contributions from each divergent adiabatic order, even the finite ones; otherwise these subtractions would *not* generally enforce covariant conservation of the renormalized stress tensor.

Then, we resort to a conformal Killing field (see Appendix B.3) of FLRW spaces, namely, the one that generates time translations:

$$\xi^\mu = a(t)\delta^\mu_0. \tag{4.77}$$

It is not difficult to verify that this field must indeed take the form (4.77) by starting from a generic (homogeneous and isotropic) timelike field $\xi^\mu = f(t)\delta^\mu_0$ and then solving the Killing equation $\pounds_\xi g_{\mu\nu} = \lambda(t) g_{\mu\nu}$ for the unknown functions f and λ. Expanding the Lie derivative:

$$\begin{aligned}\lambda g_{\mu\nu} &= \pounds_\xi g_{\mu\nu}\\ &\equiv 2\nabla_{(\mu}\xi_{\nu)}\\ &= \xi^\alpha \partial_\alpha g_{\mu\nu} + 2g_{\alpha(\mu}\partial_{\nu)}\xi^\alpha\\ &= f\partial_t g_{\mu\nu} + 2g_{0(\mu}\partial_{\nu)} f,\end{aligned} \tag{4.78}$$

for which we have the nontrivial time-time and space-space diagonal components $(0,0)$ and (i,i):

$$\begin{cases} 2\dot{f} = \lambda & \text{(4.79a)}\\ f.2a\dot{a} = \lambda a^2 & \text{(4.79b)}\end{cases} \qquad \Rightarrow \qquad \begin{cases} f = a & \text{(4.80a)}\\ \lambda = 2\dot{a} & \text{(4.80b)}\end{cases}$$

(where we have ignored a free multiplicative constant common to f and λ, setting it to 1).

Thus, using eqs. (4.74), (4.78) and (4.80b), we compute the following perfect divergence:

$$\nabla_\mu\Big(\langle T^{\mu\nu}\rangle\,\xi_\nu\Big) = \langle T^{\mu\nu}\rangle\,\nabla_{(\mu}\xi_{\nu)} = \dot{a} g_{\mu\nu}\,\langle T^{\mu\nu}\rangle = \dot{a}\,\langle T_\mu{}^\mu\rangle\,, \tag{4.81}$$

where we have explored the symmetry of the stress tensor, $T^{\mu\nu} = T^{(\mu\nu)}$, to symmetrize the derivative in the second term.

Now, integrating this divergence in a 4-volume delimited by the isotropic Cauchy surfaces Σ_{t_1} and Σ_{t_2}, we may use Gauss's theorem to obtain:

$$\begin{aligned}\int d^4x\,\sqrt{-g}\dot{a}\,\langle T_\mu{}^\mu\rangle &= \int d^4x\,\sqrt{-g}\,\nabla_\mu\Big(\langle T^{\mu\nu}\rangle\xi_\nu\Big)\\ &= \int_{\Sigma_{t_2}} d^3\mathbf{x}\sqrt{-g_{\Sigma_{t_2}}}\,\langle T^{0\nu}\rangle\xi_\nu - \int_{\Sigma_{t_1}} d^3\mathbf{x}\sqrt{-g_{\Sigma_{t_1}}}\,\langle T^{0\nu}\rangle\xi_\nu.\end{aligned} \tag{4.82}$$

In virtue of spatial homogeneity, we need not to evaluate these spatial integrals; we can simply pick any space point $\mathbf{x}$ and evaluate the integrands directly[I]. Since the

[I] Or equivalently, we can carry a (trivially homogeneous) integration in a finite coordinate volume V and then divide both sides by V. Note that homogeneity will cause the total contribution from the spatial boundaries to be null.

Jacobians $\sqrt{-g}$ and $\sqrt{-g_{\Sigma_t}}$ actually coincide for spatially flat FLRW spaces in proper-time Cartesian coordinates, $\sqrt{-g} = a^3(t) = \sqrt{-g_{\Sigma_t}}$, we obtain simply:

$$\int_{t_1}^{t_2} dt\, a^3(t)\dot{a}(t)\, \langle T_\mu^{\ \mu}\rangle(t) = a^3(t_2)\, \langle T^{00}\rangle(t_2)\, a(t_2) - a^3(t_1)\, \langle T^{00}\rangle(t_1)\, a(t_1). \tag{4.83}$$

This yields an integral expression for the energy density $\langle T^{00}\rangle = \langle T_{00}\rangle = \langle T_0^{\ 0}\rangle$ as a function of the trace $\langle T_\mu^{\ \mu}\rangle$:

$$\Big[a^4(t)\langle T_0^{\ 0}\rangle(t)\Big]_{t_1}^{t_2} = \int_{t_1}^{t_2} dt\, \dot{a}(t)a^3(t)\langle T_\mu^{\ \mu}\rangle(t). \tag{4.84}$$

Then, using eq. (4.71) for the conformally special case:

$$\begin{aligned}\int_{t_1}^{t_2} dt\, a^3(t)\dot{a}(t)\, \langle T_\mu^{\ \mu}\rangle(t) &= \frac{1}{480\pi^2}\int_{t_1}^{t_2} dt\Big[a\dot{a}\ddot{a}^2 + a^2\dot{a}a^{(4)} - 3\dot{a}^3\ddot{a} + 3a\dot{a}^2a^{(3)}\Big]\\ &= g(t_2) - g(t_1),\end{aligned} \tag{4.85}$$

where $a^{(3)} \equiv \dddot{a}$ and $a^{(4)} \equiv \ddddot{a}$.

Although it is a little intricate to manually compute a primitive $g(t)$ for the integrand in (4.85), one can easily verify that one particular solution is given by:

$$g(t) \equiv \frac{1}{480\pi^2}\Big(a^2\dot{a}\,\dddot{a} + a\dot{a}^2\ddot{a} - \tfrac{1}{2}a^2\ddot{a}^2 - \dot{a}^4\Big). \tag{4.86}$$

Equation (4.84) then entails:

$$a^4(t)\,\langle T_0^{\ 0}\rangle(t) = g(t) + E \qquad \Rightarrow \qquad \langle T_0^{\ 0}\rangle(t) = \frac{g(t)}{a^4(t)} + \frac{E}{a^4(t)}, \tag{4.87}$$

where E is simply an integration constant. From the classical behaviour of the energy density of a noninteracting massless field in a FLRW spacetime (which will be discussed in further detail in the next chapter, with emphasis on the electromagnetic field), we can immediately identify this last term as a contribution from an isotropic particle distribution, whose energy density decays as $\rho \propto a^{-4}$. Thus, we identify the remaining term as that due to vacuum energy. It reads:

$$\rho_0 = \langle 0|T_0^{\ 0}|0\rangle = \frac{1}{480\pi^2}\left[-\frac{\dot{a}^4}{a^4} + \frac{\dot{a}\,\dddot{a}}{a^2} + \frac{\dot{a}^2\ddot{a}}{a^3} - \frac{1}{2}\frac{\ddot{a}^2}{a^2}\right]. \tag{4.88}$$

In virtue of spatial isotropy, the spatial components $\langle T_i{}^j \rangle$ must be diagonal and equal (see section 5.1.1 in the next chapter for more details on this argument), which allows us to immediately compute the space-space (pressure) components $\langle T_i{}^i \rangle$ from $\langle T_\mu{}^\mu \rangle$ and $\langle T_0{}^0 \rangle$:

$$\langle T_\mu{}^\mu \rangle = \langle T_0{}^0 \rangle + 3 \langle T_i{}^i \rangle \quad \Rightarrow \quad \langle T_i{}^i \rangle = \tfrac{1}{3}\Big(\langle T_\mu{}^\mu \rangle - \langle T_0{}^0 \rangle \Big) \tag{4.89}$$

(where we are not carrying a sum in the spatial index i; its repetition just stands for a diagonal component). This gives us the vacuum pressure:

$$\begin{aligned} p_0 \equiv -\langle 0 | T_i{}^i | 0 \rangle &= \tfrac{1}{3}\Big(\rho - \langle 0 | T_\mu{}^\mu | 0 \rangle \Big) \\ &= -\frac{1}{3}\frac{1}{480\pi^2}\left[\frac{\dot{a}^4}{a^4} + 2\frac{\dot{a}\,\dddot{a}}{a^2} - 4\frac{\dot{a}^2\ddot{a}}{a^3} + \frac{3}{2}\frac{\ddot{a}^2}{a^2} + \frac{\ddddot{a}}{a} \right]. \end{aligned} \tag{4.90}$$

Having obtained these expressions for a conformally trivial case, we now take a moment to interpret them, and address a few pressing questions. What do these results tell us about the properties and dynamical behaviour of vacuum energy? Are they in any way meaningful in respect to the expansion of our own universe? Also, can they be extended beyond the conformally trivial case and into a more general range of parameters (m, ξ)? (If so, how?)

First of all, we emphasize that the relations (4.88) and (4.90) were computed for a *fixed* background metric, so that they will not be generally compatible with a *dynamical* expansion *driven solely (or mainly) by* vacuum energy. Nevertheless, they should still be useful to consistently calculate the vacuum energy in an expansion dominated by other forms of matter and energy (and eventually even compute its gravitational backreaction perturbatively).

With these caveats duely noted, we proceed to analyze the properties of the vacuum energy we have calculated. From eqs (4.88) and (4.90), we see that both ρ_0 and p_0 can have either sign, depending on the 'kinematics' of the scalar factor $a(t)$. To analyze these more concretely, we evaluate them for a few simple examples of cosmological relevance. First, let us consider a power-law expansion, $a(t) \propto t^\lambda$. In this case, it is easy to see that both ρ_0 and p_0 will decay as t^{-4} ($\propto a^{-\frac{4}{\lambda}}$); the exact expressions read:

$$\begin{cases} \rho_0 = \dfrac{\lambda^2(3 - 6\lambda + \lambda^2)}{960\pi^2} t^{-4} & \text{(4.91a)} \\ p_0 = \dfrac{\lambda(4 - 11\lambda + \frac{22}{3}\lambda^2 + \lambda^3)}{960\pi^2} t^{-4} & \text{(4.91b)} \end{cases} .$$

These can be either negative or positive, depending on the value of λ (since they are both degree 4 polynomials, they will have 4 roots where they may switch in signs). We plot them rescaled by t^4 as a function of λ:

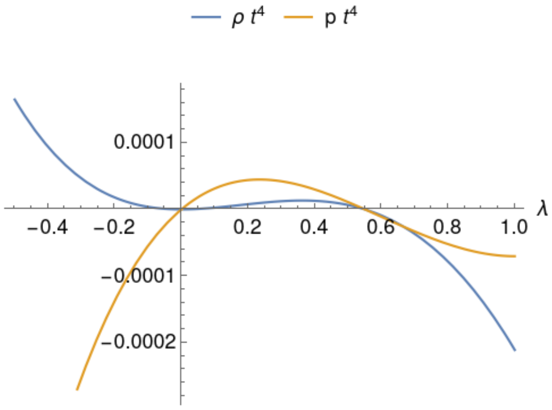

Figure 19 – Plots for the time independent scaled energy density $\rho_0 t^4$ and pressure $p_0 t^4$ arising from the vacuum renormalized values in a power-law ($a \propto t^\lambda$) FLRW universe. Note that, in the expanding region $\lambda > 0$, ρ_0 and p_0 can be both positive and negative, and they have matching signs (corresponding to a positive equation of state w_0) for $\lambda < \frac{4}{3}$; for $\lambda > \frac{4}{3}$ (region not shown in the plot) p_0 turns positive again, yielding a negative w_0.
Source: By the author.

In the context of cosmology, we define the equation of state for a given species of matter as the relation between their pressure and their energy density, in the form $p = w\rho$ (see, chapter 5 for more details). We see that the vacuum equation of state here is simply a constant, whose value depends on λ:

$$w_0 = \frac{p_0}{\rho_0} = \frac{4 - 11\lambda + \frac{22}{3}\lambda^2 + \lambda^3}{\lambda(3 - 6\lambda + \lambda^2)}. \tag{4.92}$$

Again, we emphasize this is generally not consistent with an expansion driven solely by vacuum energy. As we will see in the next chapter, the relation between w and λ that one obtains for a single species source (*i.e.*, a source with a fixed equation of state) in the Friedman Equation (5.11a) supplemented by (5.15) is (5.16):

$$w = \frac{2 - 3\lambda}{3\lambda}. \tag{4.93}$$

Two particularly important values of λ in cosmology are $2/3$ and $1/2$, corresponding to expansions driven by cold matter (“dust”) and radiation, respectively. The corresponding vacuum equation in these backgrounds would be:

$$\lambda = \frac{1}{2} \begin{cases} \rho_0(t) = \dfrac{1}{15360\pi^2}\, t^{-4} & (4.94a) \\ p_0(t) = \dfrac{1}{9216\pi^2}\, t^{-4} & (4.94b) \\ w_0(t) = \dfrac{5}{3} & (4.94c) \end{cases} \qquad \lambda = \frac{2}{3} \begin{cases} \rho_0(t) = -\dfrac{1}{3888\pi^2}\, t^{-4} & (4.95a) \\ p_0(t) = -\dfrac{1}{3888\pi^2}\, t^{-4} & (4.95b) \\ w_0(t) = 1 & (4.95c) \end{cases}$$

In both cases, we have a positive equation of state, $w_0 > 0$, although the vacuum energy density (and pressure) do alternate their sign between these two values, being positive for $\lambda = 1/2$ and negative for $\lambda = 2/3$. Curiously, the latter corresponds to the only nontrivial case $\rho_0, p_0 \neq 0$ for which both quantities coincide: $\rho_0 = p_0$.

Finally, we consider the case of an exponential expansion, $a \propto e^{Ht}$. This is relevant for a universe dominated by a form of energy which behaves like a cosmological constant, such as our current universe, dominated by Dark Energy, or many primordial inflationary scenarios. In this case we obtain simply constant energy densities and pressures, in the form:

$$\begin{cases} \rho_0(t) = \dfrac{H^4}{960\pi^2} = cte, & (4.96a) \\ p_0(t) = -\dfrac{H^4}{960\pi^2} = cte, & (4.96b) \\ w_0(t) = -1. & (4.96c) \end{cases}$$

This yields precisely a form of energy behaving like a cosmological constant!

$$\langle 0|T_{\mu\nu}|0\rangle = \Lambda g_{\mu\nu}, \qquad \Lambda = \frac{H^4}{960\pi^2}. \tag{4.97}$$

Thus, precisely for an exponential expansion, we find a form of vacuum energy which is *qualitatively* self-consistent with the expansion that it would generate. However, we emphasize that the *quantitative* consistence is still not generally satisfied. The value of Λ that is obtained in (4.97) is *generally* not the one that would produce an expansion rate H in Einstein's Equations; the latter would be proportional to H^2. Writing it as an energy contribution (i.e., as a term in $T_{\mu\nu}$, as in eq. (4.97)) in the RHS of Einstein's equations, and recovering the constants $G, \hbar, c$ in our equations, we can compare both relations between H and Λ:

$$H^2 = \frac{8\pi G}{3c^2}\Lambda; \qquad H^4 = \frac{960\pi^2 c^3}{\hbar}\Lambda, \tag{4.98}$$

If we wish to find out for which value of Λ (i.e., for which value of vacuum energy density) those would match, we square the first equation and substitute in the second, yielding:

$$\left(\frac{8\pi G}{3c^2}\right)^2\Lambda^2 = \frac{960\pi^2 c^2}{\hbar} \qquad \Rightarrow \qquad \Lambda = 45\frac{c^7}{\hbar G} = 45\rho_p, \tag{4.99}$$

which is 45 times larger than Planck energy density! Thus, we see that what would be a quantitatively self-consistent case can no longer be described in terms of classical spacetime (*for these particular values of field parameters*).

As of the last question, of whether (and how) we can extend our results beyond conformally trivial case, the answer is yes, but the calculations will be considerably more complex, and it is very difficult to avoid symbolical and numerical calculations. Although we will no longer have analytical results in these cases, there are a few characteristics that we would like to anticipate: (i) these basic features that the renormalized energy and pressure can have either sign remain valid, and, particularly, for an exponential expansion we *always* obtain a vacuum energy with an equation of state $p_0 = -\rho_0$; (ii) all the deductions we have made from eq (4.74) to (4.84) remain valid for more general m and ξ. The fundamental difference is that the trace appearing in the RHS of (4.84) *will no longer be purely anomalous* – one must generally compute the full expression of the tensor trace, which will include subtractions of 0th and 2nd adiabatic orders, as well as a contribution stemming from the *exact* field modes $|h_{\mathbf{k}}|^2$. Thus, in these cases, one must either work in very special FLRW for which known analytical solutions exist[J], or work directly with numerical ones. In the next section, we shall take the first approach and carry a detailed analysis of the renormalized stress tensor in exponentially expanding (de Sitter) spacetimes, which not only are more tractable but also happen to be a case of high interest in inflationary cosmology.

4.2.4 Renormalization in de Sitter Spacetimes: analyzing the power spectrum and the stress tensor in the Bunch-Davies vacuum

Now that we have developed the basic procedures of adiabatic subtraction and some of its applications in general FLRW spacetimes, we will specialize our approach to a more specific class of spacetimes: de Sitter spaces. De Sitter spaces are a class of curved, yet maximally symmetric spacetimes, corresponding to solutions of the Einstein Equations with a positive cosmological constant (or a spacetime homogeneous form of energy behaving like a positive cosmological constant) and no other types of matter or energy, being therefore of high interest to study inflationary scenarios dominated by this

[J] To this author's knowledge, such solutions are still quite scarce in the literature in the massive case. For solutions in the massless, *minimally* coupled case ($m, \xi = 0$), see (37). Also, there is a recent analytic treatment for massless nonconformally coupled ($m = 0$, $\xi \neq 1/6$) fields in general FLRW spaces given in (38). However, in the referred work, the authors set $m = 0$ *a priori* for the quantized field, which fails to account for a term $W_k^{(4)}$ that remains finite in the $m \to 0$ limit and thus results in a null trace anomaly in the conformally special ($\xi = 1/6$) case.

particular type of vacuum energy. These will thus provide us with a nontrivial, yet tractable backgrounds to analyze our quantized fields, and allow us to obtain more results through the procedure of adiabatic subtraction.

Then, before we proceed to our field analysis, we make a brief digression about de Sitter spaces[K]. A very convenient way to visualize these spaces is by considering a 4-dimensional hyperboloid embedded in a 5-dimensional flat Lorentzian space. If this embedding space is covered with Cartesian coordinates (T, X, Y, Z, W), such that its line element is:

$$dS^2 = dT^2 - dX^2 - dY^2 - dZ^2 - dW^2, \tag{4.100}$$

the hypersurface that represents a de Sitter space can be written by the equation:

$$T^2 - X^2 - Y^2 - Z^2 - W^2 = -H^{-2}. \tag{4.101}$$

Just like Minkowski spaces, de Sitter spaces have the maximal number of Killing fields (10, in our 4-dimensional case), and there are a large number of convenient choices of coordinates that emphasize different symmetries. Particularly, just as we can cover a portion of Minkowski spacetime (which is obviously stationary) with coordinates that give it a form of a hyperbolic FLRW space[L] with $a(t) \propto t$, we can cover half of de Sitter space with a coordinate system (t, x, y, z) that makes it look like an exponentially expanding FLRW space:

$$\begin{cases} T = H^{-1}\sinh(Ht) + \frac{1}{2}He^{Ht}(x^2+y^2+z^2) & \text{(4.102a)} \\ W = H^{-1}\cosh(Ht) - \frac{1}{2}He^{Ht}(x^2+y^2+z^2) \quad , & \text{(4.102b)} \\ X = xe^{Ht}, \qquad Y = ye^{Ht}, \qquad Z = ze^{Ht} & \text{(4.102c)} \end{cases}$$

where (t, x, y, z) all range from $-\infty$ to ∞, covering half of the hyperboloid with $T+W > 0$.

In these coordinates, the line element reads:

$$ds^2 = dt^2 - e^{2Ht}(dx^2 + dy^2 + dz^2). \tag{4.103}$$

Then, summarizing the spatial coordinates as $\mathbf{x} = (x, y, z)$, we write this compactly as:

[K] For a more detailed analysis of these spaces, their many different coordinate systems and their role in field theory, see *e.g.* section 5.2 of (25) and section 5.4 of (1).

[L] This is known in the literature as the Milne universe. See *e.g.* section 5.3 of (1).

$$ds^2 = dt^2 - e^{2Ht}d\mathbf{x}^2 = (H\eta)^{-2}\Big(d\eta^2 - d\mathbf{x}^2\Big), \tag{4.104}$$

where we have given the expression in conformal time η; in this case, it can be written as a function of t simply as:

$$\eta = \int^t e^{-Ht'}dt' = -H^{-1}e^{-Ht}. \tag{4.105}$$

At this point, we emphasize that although an *eternally* inflating universe (both to the past and to the future) that scales *exactly* exponentially can be analytically extended in a full de Sitter space, a space that is only approximately exponentially expanding for a finite time period *does not have all de Sitter symmetries* (so that it cannot be extended in a full de Sitter space) and is preferrably represented by nonstationary, exponentially expanding coordinates[M].

For this spacetime, all curvature tensors are quite simple, as they are highly constrained by symmetry. The only independent quantity is the curvature scalar $R = 12H^2 = cte$[N]. (We give a more complete account of the computation of curvature in FLRW spacetimes in section 5.1.1; in particular, the reader can easily verify that this result follows from (5.10).) The fact that H is a constant turns out to yield relatively simple dynamical equations for our scalar field (3.23):

$$\partial_t^2\phi + 3H\partial_t\phi - e^{-2Ht}\nabla^2\phi + M^2\phi = 0, \tag{4.106}$$

where we have defined a new, constant mass parameter $M^2 = m^2 + 12\xi H^2$.

We already know that a particularly convenient mode decomposition in FLRW spaces in given by (4.18). In a de Sitter space, it reads:

$$f_{\mathbf{k}}(x) = \frac{e^{i\mathbf{k}\cdot\mathbf{x}}}{\sqrt{V}}e^{-\frac{3}{2}Ht}h_{\mathbf{k}}(t). \tag{4.107}$$

Once again, this results in time-dependent harmonic oscillator (4.19) for $h_k(t)$. To solve this equation exactly in de Sitter spaces, it is convenient to perform a change in

[M] The situation is somewhat similar to the Schwarzschild spacetime, which allows for the Kruskal extension only for eternal Black Holes; for a Black Hole that forms from the collapse of matter, only a portion of an approximate Kruskal space makes physical sense (see *e.g.* chapter 6 of (5)).

[N] The remaning curvature tensors are given simply by R and appropriate combinations of $g_{\mu\nu}$. We have: $R_{\mu\nu} = 3H^2 g_{\mu\nu}$ and $R_{\mu\nu\alpha\beta} = H^2(g_{\mu\alpha}g_{\nu\beta} - g_{\nu\alpha}g_{\mu\beta})$.

variables of the form $t \to v \equiv kH^{-1}e^{-Ht} = -k\eta$ [O]. Carrying this substitution through, it is straighforward to verify that we arrive at a Bessel equation:

$$v^2 \frac{d^2 h_k}{dv^2} + v \frac{dh_k}{dv} + (v^2 - \nu^2) h_k = 0, \tag{4.108}$$

where we have defined:

$$\nu \equiv \left(\frac{9}{4} - \frac{M^2}{H^2} \right)^{12} = \left(\frac{9}{4} - \frac{m^2}{H^2} - 12\xi \right)^{\frac{1}{2}}. \tag{4.109}$$

The general solutions to these equations may then be immediatly written in terms of known special functions (see, for example (20)). A particularly convenient basis to decompose them is given by the Hankel functions $H_\nu^{(1)}$ and $H_\nu^{(2)}$:

$$h_k(t) = \sqrt{\frac{\pi}{2H}} \Big(E(k) H_\nu^{(2)}(v) + F(k) H_\nu^{(1)}(v) \Big), \tag{4.110}$$

where we have already included a normalization factor for later convenience, and $E(k), F(k)$ are numerical factors necessary to make a general linear combination of the two solutions for each value of k (we keep a k dependence here because these factors can in principle be fixed to different values for different modes h_k when we impose the adiabatic condition below) [P].

Having obtained this general solution, we would like to fix the factors $E(k)$ and $F(k)$ appropriately to obtain a subset of these exact field modes $\{f_{\mathbf{k}}\}$ *which obey the adiabatic condition (i.e. whose asymptotic behaviour corresponds to positive-frequency modes)*. Since we are dealing with a spacetime that is dynamical *at all times* (i.e. that has no asymptotically static regions) and whose dynamics is governed by a single parameter H, a seemingly natural way to investigate the adiabatic limit would be to take $H \to 0$, for which we approach a static (Minkowski) spacetime. However, it turns out that a much more convenient way to analyze this limit is simply by keeping H fixed and look at the UV ($k \to \infty$) behaviour of modes. If we take $k \gg m, H$, such that frequency is approximately just $\omega(t) \approx k/a(t)$, the adiabatic condition reads:

[O] We warn here that some authors also use the variable $u = -v$, which is an increasing function of time. This convention leads to switch of roles of the solutions $H_\nu^{(1)}$ and $H_\nu^{(2)}$ below in respect to the adiabatic condition.

[P] Bear in mind that the wave vector $\mathbf{k}$ and the (proper-time) variable t in the ODE (4.19) are independent variables, so that these factors are just constants in the field equations (4.106). Similarly, they should not be considered variables in eq (4.108), even though v was defined proportionally to k; these factors were just included in the general solutions h_k as *coefficients* of the equation solutions.

$$h_k \sim (2ke^{-Ht})^{-1/2} \exp\Big(-i\textstyle\int^t ke^{-Ht'}dt'\Big) = (2Hv)^{-1/2}e^{+iv}, \tag{4.111}$$

where we have ignored a global phase factor that emerges in the indefinite integral in the exponent.

We then wish to match the adiabatic form (4.111) with the UV-limit of (4.110). To evaluate the latter, we make use of the asymptotic form of the Hankel functions for large v (see (21), p. 920):

$$H_\alpha^{(1)}(v) \longrightarrow \sqrt{\frac{2}{\pi}}v^{-1/2}e^{i(v+\theta(\alpha))}, \qquad H_\alpha^{(2)}(v) \longrightarrow \sqrt{\frac{2}{\pi}}v^{-1/2}e^{-i(v+\theta(\alpha))}, \tag{4.112}$$

where $\theta(\alpha) = \frac{\pi}{4} + \frac{\pi\alpha}{2}$ is merely a global (spacetime-indepedent) phase factor.

Comparing these with (4.111), and noting the normalization factor that we have included in (4.110), we immediately find that, in the UV-limit:

$$E(k) \to 0, \qquad F(k) \to 1, \qquad \text{when } k \to \infty. \tag{4.113}$$

Now, to extrapolate this condition to any frequencies, we shall make use of the symmetries in de Sitter spaces. These spaces will obviously have the 6 FLRW symmetries corresponding to spatial translations and rotations, of which we have already made use in our mode decomposition. Further, we shall use a symmetry associated with a time translation, which, in our coordinate system, takes the form[Q]:

$$\begin{cases} t \to t' = t + t_0 & (4.114a) \\ \mathbf{x} \to \mathbf{x}' = \mathbf{x}e^{-Ht_0} & (4.114b) \end{cases}.$$

We also define a transformed wave vector, $\mathbf{k}' = \mathbf{k}e^{Ht_0}$, such that:

$$\frac{\mathbf{k}'}{a(t')} = \frac{\mathbf{k}}{a(t)} \qquad \text{and} \qquad \mathbf{k}' \cdot \mathbf{x}' = \mathbf{k} \cdot \mathbf{x}. \tag{4.115}$$

Then, we write the transformed modes:

$$h_{k'}(t') = \sqrt{\frac{\pi}{2H}}\Big(E(k')H_\nu^{(2)}(v) + F(k')H_\nu^{(1)}(v)\Big). \tag{4.116}$$

[Q] As with any spacetime transformations, one can either take the "active" perspective, conceiving this as an actual spacetime transformation, or the "passive" one, conceiving it merely as a change in coordinates covering spacetime.

Note that (4.116) only differ from (4.110) in the factors $E(k), F(k) \longrightarrow E(k'), F(k')$. We then argue that a priviledged family of modes in de Sitter spacetime, in respect to which we should define a vacuum state, should be one that is invariant under these symmetries as well. That is, it should obey:

$$f_{k'}(x') = f_k(x) \qquad \Rightarrow \qquad h_{k'}(t') = h_k(t). \tag{4.117}$$

In this case, the asymptotic form (4.113) will imply that:

$$E(k) = 0, \qquad F(k) = 1, \qquad \forall k. \tag{4.118}$$

We then have the solutions that match the adiabatic condition (the asymptotically "positive-frequency") solutions:

$$h_k(t) = \sqrt{\frac{\pi}{2H}} H_\nu^{(1)}(v) = \sqrt{\frac{\pi}{2H}} H_\nu^{(1)}\left(\frac{k}{H} e^{-Ht}\right) \tag{4.119}$$

$$\Rightarrow \qquad f_{\mathbf{k}}(x) = \sqrt{\frac{\pi}{2H}} \frac{e^{i\mathbf{k}\,\cdot\,\mathbf{x}}}{\sqrt{2(2\pi)^3 e^{3Ht}}} H_\nu^{(1)}\left(\frac{k}{H} e^{-Ht}\right). \tag{4.120}$$

Then, quantizing the field through the usual mode expansions (3.31) in $\{f_{\mathbf{k}}, f^*_{\mathbf{k}}\}$, we can define a vacuum state $|0\rangle$ associated to them. This is known in the literature as the *Bunch-Davies vacuum.* Besides the usual UV divergences, ubiquitously present in QFT, this vacuum state is known to suffer from infrared divergences in the field amplitudes $\langle\phi^2\rangle$ and the stress tensor $\langle T_{\mu\nu}\rangle$ in the minimally coupled, massless case. These divergences emerge due to the higher-order singularities in the Hankel functions $H_\nu(x)$ as $x \to 0$ for $\nu \geq \frac{3}{2}$; for the same reason, it can be troublesome to evaluate parameters for which $M^2 < 0$ [R] with this vacuum state. (Surprisingly, it turns out that the adiabatic subtraction procedure, designed to eliminate UV divergences, also cancels the IR divergences in this $M^2 = 0$ $(\nu = 3/2)$ case[S]; we note, however, that such subtracted divergences can still be a delicate matter in a numerical treatment, both in the IR and in the UV.) Notwithstanding, the modes (4.120) and the Bunch-Davies vacuum will constitute the basis of our analysis of field theory in de Sitter spaces.

[R] One could argue that a $M^2 < 0$ case is in itself pathological, due to vacuum instabilities; however, this will actually be a relevant regime for stable interacting theories with local maxima in their potentials, as we shall see in the next chapter.

[S] See section 2.10 (2). For a more general treatment of vacuum states and IR divergences in FLRW spaces, including the power-law and exponential cases, see (37). See also (39, 40) for a rigorous and detailed account of vacuum states in de Sitter spacetime based in its symmetry groups.

In what follows, we will consider a scalar field with a Lagrangian (3.23) and carry a numerical analysis of two crucial renormalized observables, namely, the quadratic field amplitudes $\langle\phi^2\rangle$ and the stress tensor varying the parameters m and ξ through a region of parameters.

A good starting point for our renormalization analysis are the expectation values of the field amplitudes $\langle 0|\phi^2(x)|0\rangle$. This observable has a sufficient simple form for us to take a closer look at its spectral expansion and consider adiabatic subtractions to different orders. First, let us consider its formal, unsubtracted expansion:

$$\langle 0|\phi^2(x)|0\rangle = \frac{1}{4\pi^2 a^3(t)}\int_0^\infty dk\, k^2|h_k(t)|^2 \equiv \int_0^\infty \frac{dk}{k}\mathcal{P}_0(k,t), \tag{4.121}$$

where we have defined the *power spectrum* $\mathcal{P}(k,t)$ (and we use the subscript 0 in (4.121) to emphasize that it refers to the unsubtracted value). In terms of our field modes, it reads:

$$\mathcal{P}_0(k,t) = \frac{k^3|h_k(t)|^2}{4\pi^2 a^3(t)} = \frac{k^3 e^{-3Ht}}{8\pi H}\left|H_\nu^{(1)}\left(\frac{k}{H}e^{-Ht}\right)\right|^2. \tag{4.122}$$

In terms of the power spectrum, it is easy to see that the integral (4.121) will diverge in the IR ($k \to 0$) whenever $\mathcal{P}$ is nonvanishing in this lower limit (and, generally, it will diverge in the UV). We plot the form of this power spectrum at a fixed time $t = 0$ for a massive, minimally coupled field (Figure 20) :

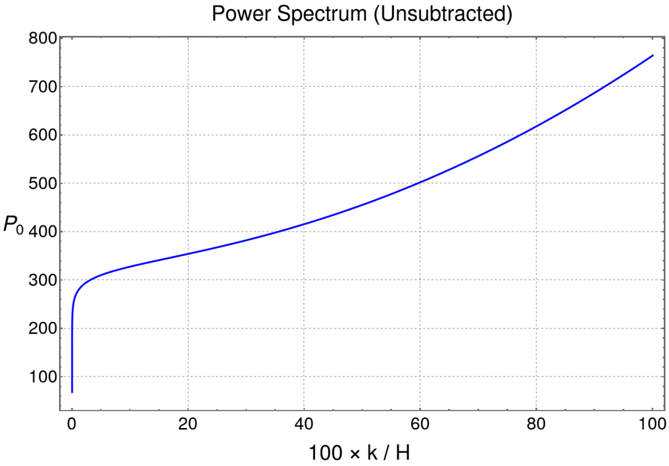

Figure 20 – Unsubtracted power spectrum for a minimally coupled ($\xi = 0$) field with mass $m^2 = 0.1H^2$. In this case $\nu = [\frac{9}{4} - \frac{1}{10}]^{\frac{1}{2}} < \frac{3}{2}$, such that $\mathcal{P}_0$ is UV-divergent but it vanishes as $k \to 0$.
Source: By the author.

This can be seen to yield a divergent expectation value for $\langle\phi^2\rangle$. According to our adiabatic subtraction prescription, it should be subtracted *only* up to second adiabatic order to render the corresponding finite physical result. By carrying this adiabatic subtraction, we obtain:

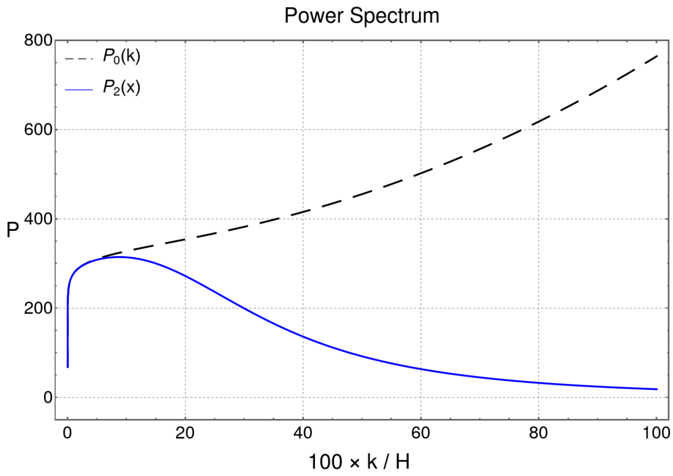

Figure 21 – Unsubtracted (black dashed line) and 2nd-order-subtracted (Blue filled line) Power Spectra for $\xi = 0$ and $m^2 = 0.1H^2$. While the former is UV-divergent, the latter yields an integrable spectrum, which is associated with the renormalized observable $\langle\phi^2(x)\rangle_{phys}$.
Source: By the author.

We shall analyze the power spectrum in more detail in chapter 5, when we discuss the potential role of such vacuum fluctuations in the fluctuation spectrum of the Cosmic Microwave Background (CMB) observed today. For completeness, we also show at this point how the Power Spectrum would look like when subtracted up to 4th order, which will be the one relevant in computing the stress tensor (Figure 22):

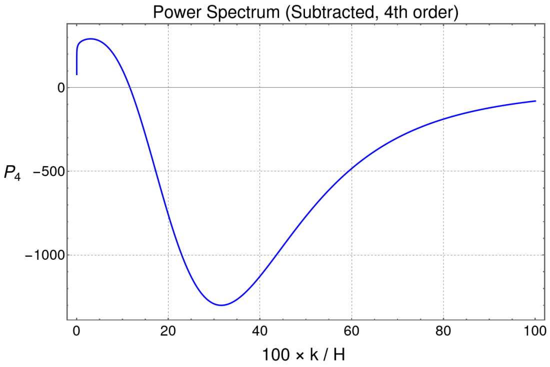

Figure 22 – Power Spectrum $\mathcal{P}_4$, subtracted up to 4th Adiabatic order for $\xi = 0$ and $m^2 = 0.1H^2$. For these particular parameters, it is only at this order that the loss of positive-definiteness manifests. However, we note once again that this is a general feature of renormalization, and it can manifest in any subtracted orders.
Source: By the author.

Finally, we show the behaviour of the unsubtracted and the subtracted power spectra covering a comprehensive range of parameters:

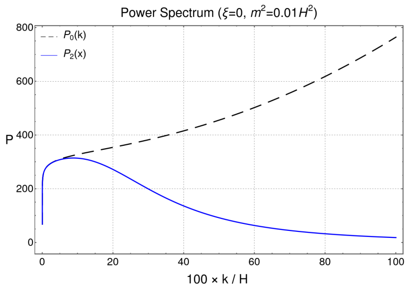

Figure 23 – $\xi = 0, m^2 = 0.01H^2$. Source: By the author.

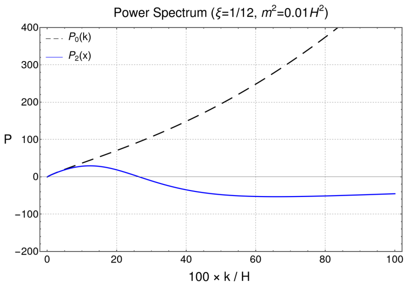

Figure 24 – $\xi = 1/12, m^2 = 0.01H^2$. Source: By the author.

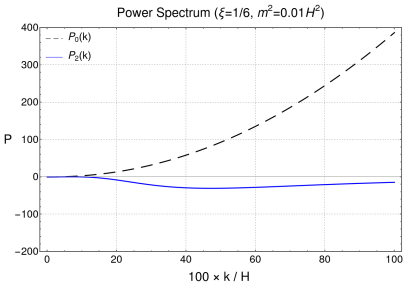

Figure 25 – $\xi = 1/6, m^2 = 0.01H^2$. Source: By the author.

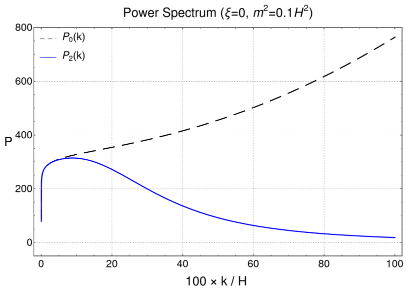

Figure 26 – $\xi = 0, m^2 = 0.1H^2$. Source: By the author.

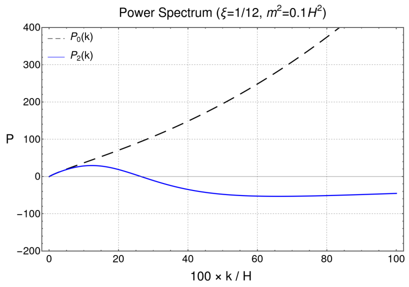

Figure 27 – $\xi = 1/12, m^2 = 0.1H^2$. Source: By the author.

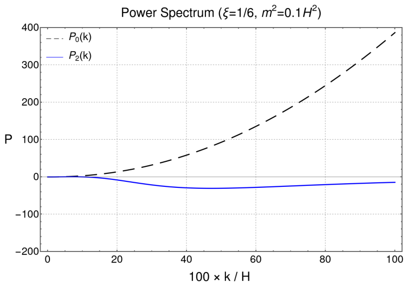

Figure 28 – $\xi = 1/6, m^2 = 0.1H^2$. By the author.

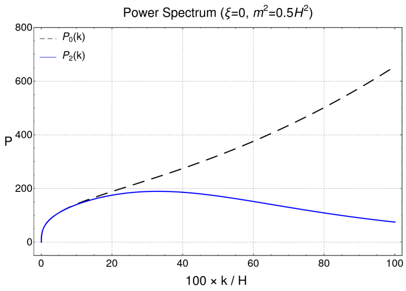

Figure 29 – $\xi = 0, m^2 = 0.5H^2$. Source: By the author.

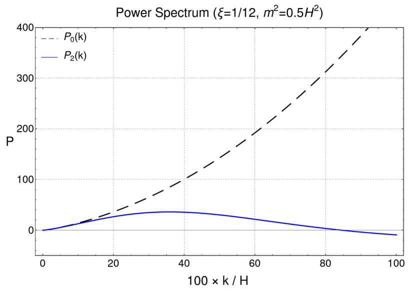

Figure 30 – $\xi = 1/12, m^2 = 0.5H^2$. Source: By the author.

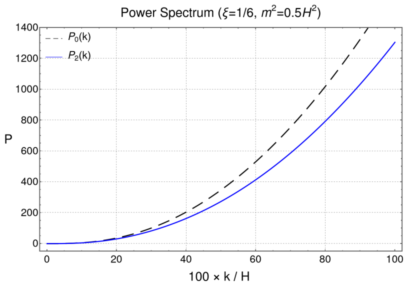

Figure 31 – $\xi = 1/6, m^2 = 0.5H^2$. Source: By the author.

For the parameters that yield $0 < M^2 < H^2$, we will have well-behaved spectra in both the IR and the UV. For $M^2 < 0$, however (as we see in Figure 33), we will have IR divergences (even if $m^2 > 0$, for which ω_k^{-1} remains bounded) even in the subtracted spectrum.

Having performed the apropriate adiabatic subtractions for an observable of interest, we must then evaluate the corresponding spectral integral to obtain its renormalized value in position space. Particularly, we are interested in the stress tensor $\langle T_{\mu\nu}(x)\rangle_{phys}$. As in the

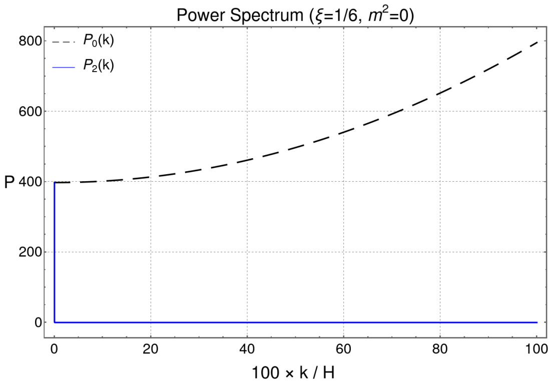

Figure 32 – $\xi = 1/6, m^2 = 0$. In this conformally trivial regime, all contributions come from the zeroth order term for any finite k, so the subtracted terms yield a trivial spectrum.
Source: By the author.

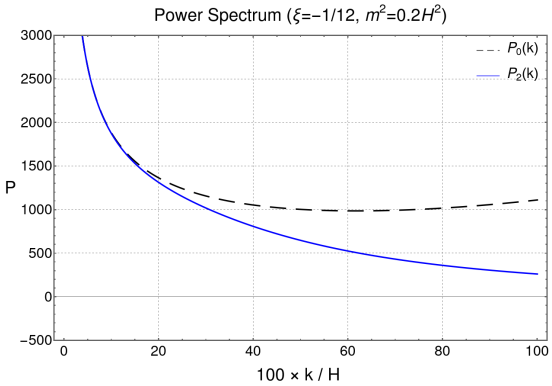

Figure 33 – $\xi = -1/12, m^2 = 0.2H^2$. For this choice of parameters, although m^2 is positive, the effective mass M^2 is negative, so that $P(k \to 0)$ is finite and there will be IR divergences
Source: By the author.

previous section, we shall obtain it from its renormalized trace, which will generally have more than just an anomalous contribution. In fact, as this trace is spatially homogeneous and it scales quite simply with time in a de Sitter space (as *all* derivatives of $a(t)$ will have a time dependence proportional to e^{Ht}), we can obtain meaningful information from it by analyzing it at a fixed event (which, for convenience, we shall fix at the origin of our coordinate system). A numerical analysis for a sample of well-behaved parameters (m, ξ) then yields:

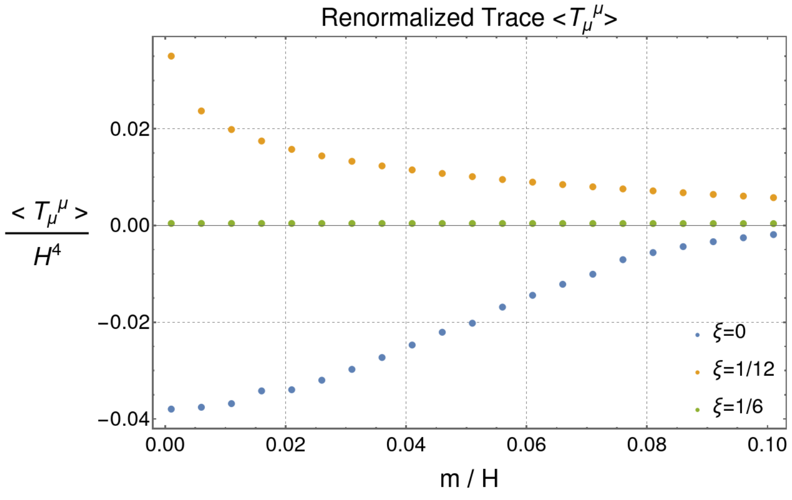

Figure 34 – Renormalized value of the stress tensor trace $\langle T_\mu{}^\mu \rangle$ at $t = 0$ normalized by H^4. In the conformally coupled case $\xi = 1/6$, its dominant contribution comes from the conformal anomaly (for this range of masses), which turns out to be relatively small compared to its nonanomalous contributions for $\xi \not\approx 1/6$.
Source: By the author.

This sample already reveals to us a number of features of the renormalized trace in de Sitter spaces. First, it can have either sign depending on both of the parameter values;

a particular consequence of this is that there will be a 1-dimensional region in the (ξ, m) plane for which $\langle T_\mu{}^\mu \rangle$ will vanish. We also stress that, although in Figure 34 (36), the trace may look vanishing in the conformally coupled case, it is actually slightly positive. In fact zooming in this plot a little (see Figure 35), we can verify that it numerically agrees with our analytical result (4.71) as $m \to 0$:

$$\frac{\langle T_\mu{}^\mu \rangle}{H^4} = \frac{1}{240\pi^2} \approx 0.00042\,. \tag{4.123}$$

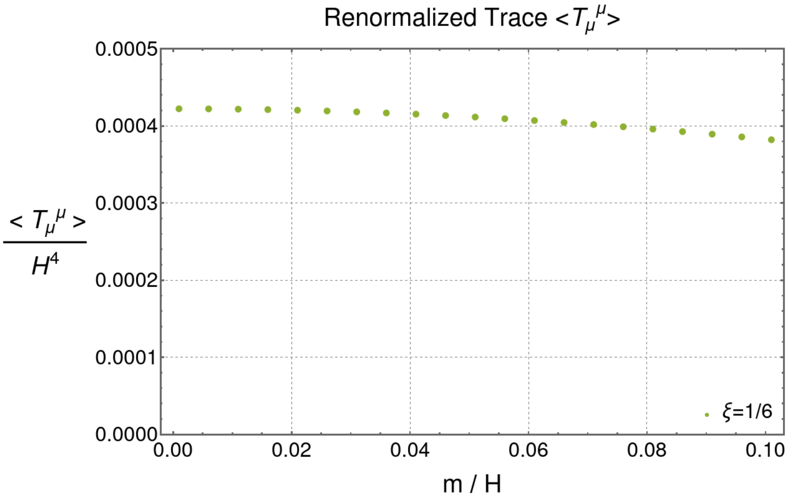

Figure 35 – Renormalized value of the stress tensor trace $\langle T_\mu{}^\mu \rangle$ at $t = 0$, normalized by H^4, for the conformally coupled case $\xi = 1/6$. In this regime, its dominant contribution comes from the conformal anomaly (for this range of masses); as $m \to 0$, one can verify its numerical agreement with (4.71).
Source: By the author.

For completeness, we plot a similar graphic to 34, with a few more values of parameters, for which we can see the trace actually cross the axis:

Note also that the sign of the trace will be determinant for the sign of the energy density, as we see in eq (4.84). By computing numerical integrals both in the spectrum and in time (for the latter we start at $t = 0$ and go through a few e-folding periods, H^{-1}), we may obtain the renormalized values of energy density and pressure, which can in principle be a function of time. We display the typical behaviour for $\rho(t)$ and $p(t)$, exemplified in a minimally coupled massive ($m^2 = 0.1H^2$) case:

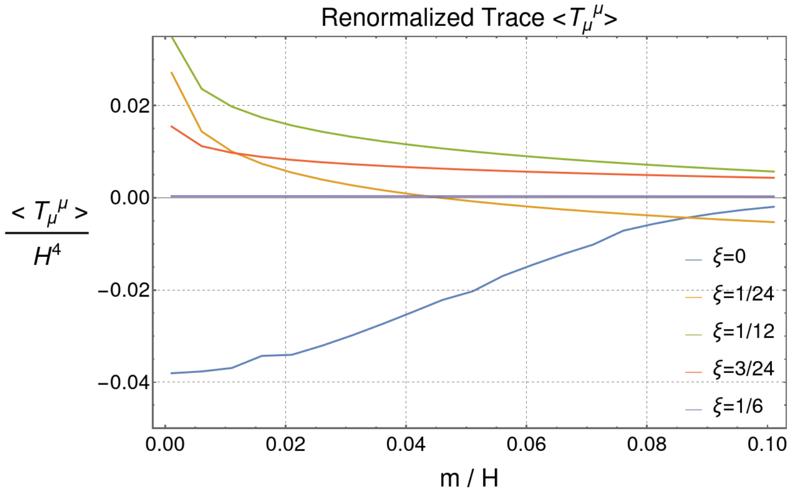

Figure 36 – Renormalized value of the stress tensor trace $\langle T_\mu{}^\mu \rangle$ at $t = 0$ normalized by H^4, plotted in lines for better visualization of the crossing values. In this sample, one can see that, for $\xi = 1/24$, $\langle T_\mu{}^\mu \rangle$ actually crosses the axis near $m = 0.045H$.
Source: By the author.

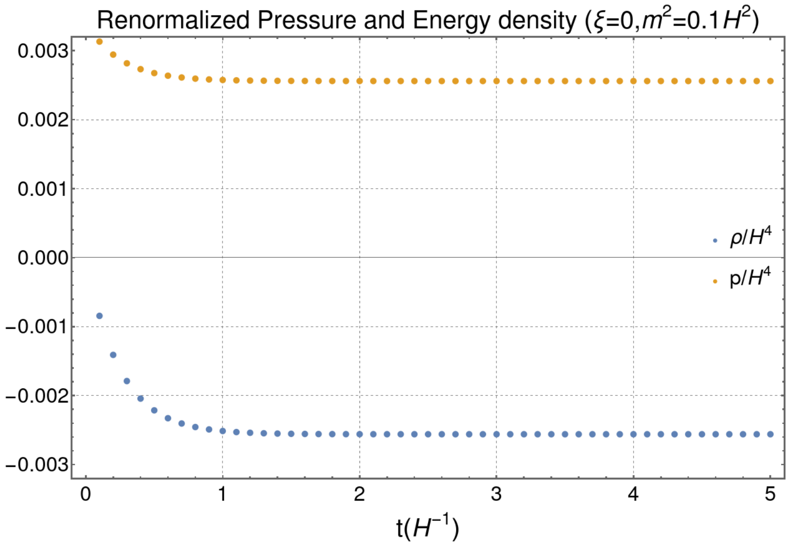

Figure 37 – Renormalized energy density and pressure as functions of time, with initial condition $\rho_0 = 0$. After a few e-foldings, these quantities evolve to constant values $\rho_{eq} \approx 0.0025H^4$ and $p_{eq} \approx -\rho_{eq}$.
Source: By the author.

Note that the initial condition imposed in our temporal integral was $\rho(t = 0) = 0$, as we carried a definite integral starting at $t = 0$. Then, after few e-foldings, one sees that energy and pressure quickly evolve to constant equilibrium values ρ_{eq} and $p_{eq} \simeq -\rho_{eq}$, obeying the equation of state that is (qualitatively) self-consistent with de Sitter spaces $(\langle T_{\mu\nu} \rangle = \Lambda g_{\mu\nu})$. Moreover when we analyze the difference term:

$$\delta\rho(t) \equiv \rho(t) - \rho_e q, \tag{4.124}$$

we find that it decays precisely as $\delta\rho(t) \propto e^{-4Ht} = a^{-4}(t)$. Then, in eq (4.87) we immediately identify this transient term as the decaying integration constant that was attributed to particle terms, rather than vacuum energy[T], so we identify the renormalized vacuum energy (pressure) as $\rho_0 = \rho_{eq}$ $(p_0 = p_{eq})$.

In this particular case, we have found $\rho_0 < 0$ and $p_0 > 0$, the signs can be reversed, depending on the values of m and ξ. In fact, carrying the same analysis for the conformally trivial case, we recover precisely our analytical results (4.96) (see Figure 38):

$$\frac{\rho_{eq}}{H^4} \simeq -\frac{p_{eq}}{H^4} \approx \frac{1}{960\pi^2} \simeq 0.00011\,. \tag{4.125}$$

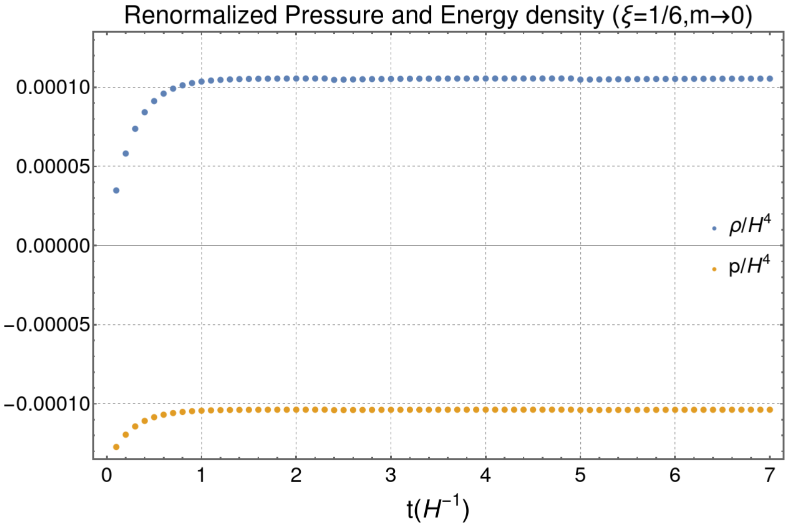

Figure 38 – Renormalized energy density and pressure as functions of time in the conformally trivial case.
Source: By the author.

Applying the same procedure for various parameter values within a well-behaved range (for which we manage to achieve numerical convergence), we find by inspection that, indeed, the equation of state for our renormalized vacuum energy is always of the form $p_0 \simeq -\rho_0$ in our de Sitter spacetimes, yielding a stress tensor in the form of a cosmological constant $\langle T_{\mu\nu}\rangle = \Lambda g_{\mu\nu}$. This is actually not surprising, as we have built the Bunch-Davies vacuum to be invariant under the de Sitter symmetries, and, in this maximally symmetrical spacetime, the only possibility for a symmetric rank (0,2) tensor built only from geometrical quantities is $\langle T_{\mu\nu}\rangle \propto g_{\mu\nu}$. Nonetheless, the present analysis has allowed us to explicitly compute the renormalized values of vacuum energy densities and pressures, for which we not only verify this self-consistency geometrical condition

[T] The same interpretation could be attained here, as the particle energy density should decay as $a^{-4}(t)$, provided that one generally considers particles with both positive and negative energies, as $\delta\rho$ can have either sign.

to be satisfied, but also obtain specific values for ρ_0 and p_0 for sufficiently well-behaved parameters.

Then, to conclude this section, we show our results for the renormalized vacuum energy ρ_0 obtained by this procedure for parameters ranging in the intervals $0 < \xi \leq 1/6$ and $0 < m^2 \leq 0.1H^2$:

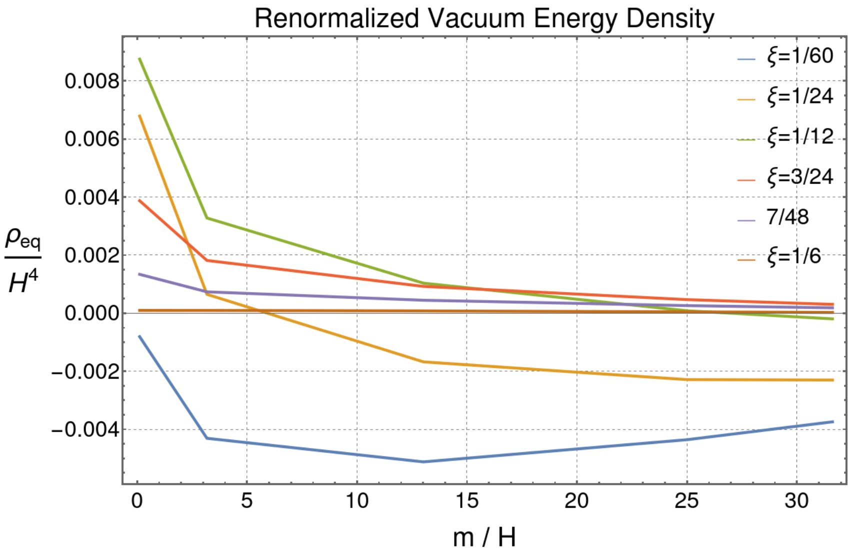

Figure 39 – Renormalized energy density for various values of m and ξ within well-behaved intervals.
Source: By the author.

As we have repeatedly remarked, this renormalized vacuum energy can be found to bear either sign as we sweep the parameters, and, particularly there will be a 1-dimensional region in the (m, ξ) plane for which it will be trivially null; again, we stress that the conformally coupled case actually lies slightly above the axis. We also note that, for $m \approx 0$, the values of ξ that approach $1/6$ successively approach the conformally anomalous vacuum energy for $\xi > 1/12$.

Finally, we note that, as the vacuum energy densities ρ_0 are significantly higher (for a fixed value of H) in the conformally nontrivial cases, or equivalently, the ratio H^4/ρ_0 are significantly lower, one finds in eq (4.98) that the self-consistency values of Λ will be smaller. If we denote the conformally trivial vacuum energy as Λ_t and a conformally nontrivial value as $\Lambda = \alpha\Lambda_t$, we find that the corresponding self-consistent values of H and Λ would yield (see eq 4.99):

$$\Lambda = \alpha^{-1}\Lambda_t = \frac{45}{\alpha}\rho_p, \tag{4.126}$$

which can yield sub-Planckian self-consistent values of Λ for a sufficiently high α. The energy densities found here, however, although significantly higher than Λ_t, have only $|\alpha| \lesssim \mathcal{O}(10^2)$, which would still correspond to a Planckian self-consistency regime.

4.3 Path integrals, effective action and Renormalization of Gravitational Parameters

Now that we have had a first operational contact with the subject of renormalization, in the concrete example of adiabatic subtraction, we would like to better understand it conceptually, and take a glimpse on its links with the wider scheme of renormalization of geometrical parameters in a gravitational context.

To achieve that, we start by giving a brief presentation of the Schwinger Action Principle, which will allow us to construct the effective action W. This action principle is intimately related to the path-integral formulation of quantum mechanics, which not only gives a novel conceptual perspective to the theory but also provides us with a very powerful arsenal to operate with the effective action. Not surprisingly, W will present divergences in field theories; to properly handle them and put them in a renormalizable form, we shall write an asymptotic expansion for W in which we can isolate the divergencies in a finite number of terms, and eventually subtract them from the matter action, reabsorbing them in the definition of geometrical parameters.

In Classical Mechanics (be it particle mechanics or classical field theory), one could derive the dynamics from an extreme action principle. This principle could be summarized as "the path that a system will classically follow to go from a configuration q_1 at time t_1 (at a Cauchy surface Σ_1) to a configuration q_2 at a time t_2 (at a Cauchy surface Σ_2) will be that which extremizes the action functional $S[q(t)]$". In Quantum Mechanics, where a precise determination of a classical path for a system is forbidden, one can see this principle in a new light. In Feynman's formulation citeFeynmanA,FeynmanB, one can state a quantum version of the action principle as follows: "the *conditional probability amplitude* that a system starting at a configuration $|q'\rangle$ at a time t_1 (a Cauchy surface Σ_1) will be measured at a configuration $|q''\rangle$[U] at a time t_2 (a Cauchy surface Σ_2) *having passed through a classical path*[V] $q'(t)$ will be proportional to $e^{iS[q'(t)]}$, being $S[q'(t)]$ the classical action associated with that path". We summarize in a condensed notation:

$$\langle q'', t_2 | q', t_1 \rangle \, [q'(t)] \propto e^{\frac{i}{\hbar} S[q'(t)]}, \tag{4.127}$$

where we temporarily recovered Planck's constant ($\hbar \neq 1$) in order to make a clearer picture of the classical limit ahead. (We also note that these probabilities will be generally

[U] Here, we avoid the most obvious notation $|q_1'\rangle$ and $|q_2'\rangle$ for those configurations to avoid an ambiguity in the following discussion, where we shall consider *different observables* q_1 and q_2 to be evaluated at times t_1 and t_2; in this case q_1' and q_2' will denote *eigenvalues of distinct operators*, whereas, for instance, q_1' and q_1'' denotes *distinct eigenvalues of the same operator.*

[V] On first sight, this may sound contradictory to what we just said above, that the *determination* of a classical path is forbidden, but it is not. We are not *determining* a classical path, just associating conditional probability amplitudes to it. This is the same as, *e.g.,* associating amplitudes from the 'left' and 'right' paths in a double slit experiment; one may then appropriately add these amplitudes to obtain an interference pattern.

nonnormalizable, and one must use sophisticated functional integral techniques to handle them.)

With this physical postulate, one can in principle work out any dynamical predictions of a quantum theory – which can be put in the form of transition amplitudes $\langle q'', t_2|q', t_1\rangle$ – by summing (integrating) the *conditional amplitudes* (4.127) over all classical paths $q'(t)$, *i.e., by calculating a path integral.* These predictions turn out to be equivalent to those obtained by the known quantum dynamical equations obtained through canonical quantization; then, conversely, one can deduce these equations from Feynman's basic postulate.

More generally, one could measure any complete set of commuting observables (C.S.C.O.) to maximally determine a physical state, rather than just its "position" configurations $|q'\rangle$; then, denoting eigenstates of two arbitrary C.S.C.O.'s A and B by $|a'\rangle$ and $|b'\rangle$, respectively, the most general probabilities that we could measure are of the form $\langle b'|a'\rangle$. Usually, as we will be interested in determining transition amplitudes for arbitrary C.S.C.O.'s at times t_1 and t_2 (Cauchy surfaces Σ_1 and Σ_2), we will correspondingly denote these arbitrary observables as q_1 and q_2 (or ζ_1 and ζ_2) and the amplitudes $\langle q_2', t_2|q_1', t_1\rangle$ (or $\langle \zeta_2', \Sigma_2|\zeta_1', \Sigma_1\rangle$).

Additionally, in this formulation, the classical limit of the theory becomes conceptually quite obvious. Whenever a physical system has an action $S[t]$ that is very large with respect to Planck's constant, such that $S[q(t)]/\hbar \gg 1$ and it varies by a great amount for small path variations, then interference from nearby paths will cancel out almost everywhere, and *nonneglegible probabilities will arise only from the paths around which the action varies the least, i.e., around stationary paths, which extremize the action.* Then, in this limit, the probability that a system goes very nearly around a path that extremizes the action becomes practically unity, so that one recovers the system's classical paths as given by the stationary action principle.

Besides, this quantization formalism extends very directly to interacting fields, whereas our approach in chapter 3 was restricted to noninteracting fields, which could be conveniently put in a form of an infinite collection of decoupled harmonic oscillators. However, path integrals are generally very complicated to calculate. The noninteracting case is one of the few where calculations can be rendered in a computable form, using (infinite-dimensional) Gaussian integrals.

An equivalent and intimately related way to define this manifestly covariant approach to quantum theory is through the Schwinger Action Principle[W]. In this formulation, one works directly with transition amplitudes and their variations in terms of an action

[W] For the original works on Schwinger's formulation, see the seminal papers citeschwingerI,schwingerII. For a more pedagogical and thorough textbook introduction, see citetoms.

operator: the *effective action*. Thus, it will be more convenient for us to start with this formalism, as it will allow us to directly *define* the effective action as a starting point and then *derive* its relation to path integrals.

Before presenting it in the context of field theory, it is constructive to briefly illustrate our considerations in the simpler context of ordinary quantum mechanics. A first point that we stress here is that we are describing the possible *physical states* (configurations) in our system by *complete sets of eigenvalues of time-dependent observables*. Thus, when we write $|a', t\rangle$ as an eigenstate of the C.S.C.O. $A(t)$, we mean that:

$$|a', t\rangle : \; A(t)\, |a', t\rangle = a'\, |a', t\rangle \, , \tag{4.128}$$

where the eigenvalue a' is time-independent. This representation has the potentially confusing implication that *our basis state vectors will be generally time-dependent in the Heinsenberg picture and time-independent in the Schrödinger picture*. Indeed, consider for instance a free nonrelativistic particle moving in 1 spatial dimension and the C.S.C.O. given by the position operator $x(t)$: in the Schrödinger picture, x is time independent, thus so will be its eigenvectors $|x'\rangle$ (up to an arbitrary phase factor, which we omit), since:

$$x\, |x', t\rangle = x'\, |x', t\rangle \, , \forall t \quad \Rightarrow \quad |x', t_1\rangle = |x', t_2\rangle \, ; \tag{4.129}$$

in the Heisenberg picture, on the other hand, x is time-dependent, such that $x(t_1) \not\equiv x(t_2)$, which implies that:

$$x(t)\, |x', t\rangle = x'\, |x', t\rangle \, , \forall t \quad \Rightarrow \quad |x', t_1\rangle \neq |x', t_2\rangle \, . \tag{4.130}$$

Note, further, that, if Schrödinger ordinary state vectors evolve by the evolution operator U ($|\psi(t)\rangle = U(t)\psi_0$), and thus Heisenberg operators evolve as $U^\dagger(t) A_0 U(t)$, then our *Heisenberg basis vectors* will evolve by the *inverse* evolution operator $U^\dagger$, in order to satisfy (4.130) at all times.

Having clarified our basic notation, we proceed to discuss the dynamics. Let q_1 and q_2 be two C.S.C.O.'s in our theory, so that our fundamental physical description may be given by the transition rates:

$$\langle q_2', t_2 | q_1', t_1 \rangle \, . \tag{4.131}$$

Then, the Schwinger Action Principle ascertains that variations on those transitions will take the form:

$$\delta\langle q_2', t_2|q_1', t_1\rangle = i\,\langle q_2', t_2|\,\delta S\,|q_1', t_1\rangle\,, \tag{4.132}$$

where S is the action functional of the theory. Just like (2.1) or (2.13), it is to be regarded as a function of the basic dynamic variables; then, in the quantum case, it will be an operator. We then define the *effective action* W by:

$$\langle q_2', t_2|q_1', t_1\rangle = e^{iW}. \tag{4.133}$$

From (4.132) and (4.133), we find that variations in W read:

$$\delta W = \frac{\langle q_2', t_2|\,\delta S\,|q_1', t_1\rangle}{\langle q_2', t_2|q_1', t_1\rangle}. \tag{4.134}$$

In relativistic field theory, we shall have a very similar situation. The most relevant distinction to our discussion so far is that we must trade the simple notion of a time instant for that of a Cauchy surface, so our basic configurations are correspondingly modified $|q', t\rangle \to |\zeta', \Sigma\rangle$, and our transition amplitudes read:

$$\langle\zeta_2', \Sigma_2|\zeta_1', \Sigma_1\rangle \qquad \Rightarrow \qquad \delta W = \frac{\langle\zeta_2', \Sigma_2|\,\delta S\,|\zeta_1', \Sigma_1\rangle}{\langle\zeta_2', \Sigma_2|\zeta_1', \Sigma_1\rangle}. \tag{4.135}$$

Now, these variations in the action could come from different sources. The most obvious one is a change in the dynamical variables, of which it is a direct function. But one could also vary the basic *parameters* of the theory (such as masses and coupling constants, or even adding external sources) or Cauchy-Surface boundaries of the action integral (2.13).

A convenient way for us to analyze variations in the action is by modifying it through the addition of an external classical source term. In the case of a single scalar field, we introduce a scalar source J, modifying the action S as follows:

$$S[\phi] \longrightarrow S_J[\phi] = S[\phi] + \int d\mu_g(x) J(x)\phi(x). \tag{4.136}$$

Thus, the field equations are correspondingly modified to:

$$\frac{\delta S_J}{\delta\phi(x)} = \frac{\delta S}{\delta\phi(x)} + J(x) = 0. \tag{4.137}$$

Considering the original field equations (3.25), this yields:

$$\Big[\Box_x + m^2 + \xi R(x)\Big]\phi(x) = -J(x). \tag{4.138}$$

We then consider the basic functional Z of our theory, which takes the form of transition amplitudes:

$$Z[J] \equiv \langle \zeta_2', \Sigma_2 | \zeta_1', \Sigma_1 \rangle \, [J] = \langle 2|1 \rangle \, [J], \tag{4.139}$$

where we have once a simplified notation in the last equality, denoting our states as $|1\rangle\,, |2\rangle$. Note that Z relates to the effective action W simply as:

$$W[J] = -i \ln(Z[J]) \tag{4.140}$$

Once again, the Schwinger Action Principle states that *general* variations in this modified theory will take the form:

$$\delta Z[J] = i \, \langle 2| \, \delta S_J \, |1\rangle \, . \tag{4.141}$$

Particularly, if we vary *only* the source J, leaving the remaining parameters, the dynamical variables and the integration boundaries fixed, we obtain a simple variation in the form:

$$\delta Z[J] = i \int d\mu_g(x) \Big(\langle 2|\phi(x)|1\rangle[J] \Big) \delta J(x), \tag{4.142}$$

or, using functional derivatives:

$$\frac{\delta Z[J]}{\delta J(x)} = i \, \langle 2|\phi(x)|1\rangle[J]. \tag{4.143}$$

This analysis can be carried on further, and we can analyze second variations of Z with respect to J – or, equivalently, variations of (4.143) with respect to it:

$$\delta\left(\frac{\delta Z[J]}{\delta J(x)}\right) = i\delta\Big(\langle 2|\phi(x)|1\rangle[J]\Big). \tag{4.144}$$

To evaluate the RHS of (4.144) it will be convenient to consider an intermediate Cauchy surface Σ between Σ_1 and Σ_2, containing the event x, and decompose the source variation in two parts: $\delta J(x') = \delta J_1(x') + \delta J_2(x')$, such that δJ_1 vanishes identically to the future of Σ, $I^+(\Sigma)$ and δJ_2 vanishes identically to its past, $I^-(\Sigma)$. Correspondingly, we split the total variation in the form:

$$\delta \langle 2|\phi(x)|1\rangle = \delta_2\langle 2|\phi(x)|1\rangle + \delta_1\langle 2|\phi(x)|1\rangle\,. \tag{4.145}$$

The trick now is to write the completeness relation with the eigenstates of an intermediate C.S.C.O. ζ in Σ, $|\zeta',\Sigma\rangle$, to conveniently evaluate each of these variations in terms of (4.142). For example, for δ_1, we write:

$$\begin{aligned}\delta_1 \langle 2|\phi(x)|1\rangle &= \sum_{\zeta'} \delta_1\Big(\langle 2|\phi(x)|\zeta'\rangle\langle\zeta'|1\rangle\Big)\\ &= \sum_{\zeta'}\Big(\delta_1\langle 2|\phi(x)|\zeta'\rangle\Big)\langle\zeta'|1\rangle + \langle 2|\phi(x)|\zeta'\rangle\Big(\delta_1\langle\zeta'|1\rangle\Big).\end{aligned} \tag{4.146}$$

However, since δ_1 yields null variations between Σ and Σ_2, only the second term will be nonvanishing. Thus:

$$\begin{aligned}\delta_1 \langle 2|\phi(x)|1\rangle &= \sum_{\zeta'} \langle 2|\phi(x)|\zeta'\rangle\Big(\delta_1\langle\zeta'|1\rangle\Big)\\ &= i\int d\mu_g(x')\delta J_1(x')\sum_{\zeta'}\langle 2|\phi(x)|\zeta'\rangle\langle\zeta'|\phi(x')|1\rangle\\ &= i\int d\mu_g(x')\delta J_1(x')\,\langle 2|\phi(x)\phi(x')|1\rangle\end{aligned} \tag{4.147}$$

(where $\delta J_1(x)$ will vanish identically in $I^+(\Sigma)$).

Similarly, we have the variation δ_2:

$$\begin{aligned}\delta_2 \langle 2|\phi(x)|1\rangle &= \sum_{\zeta'} \delta_2\Big(\langle 2|\zeta'\rangle\langle\zeta'|\phi(x)|1\rangle\Big)\\ &= \sum_{\zeta'}\Big(\delta_2\langle 2|\zeta'\rangle\Big)\langle\zeta'|\phi(x)|1\rangle\\ &= i\int d\mu_g(x')\delta J_2(x')\,\langle 2|\phi(x')\phi(x)|1\rangle\end{aligned} \tag{4.148}$$

(where $\delta J_2(x)$ will vanish identically in $I^-(\Sigma)$).

Then, adding the two variations, we obtain:

$$\begin{aligned}\delta \langle 2|\phi(x)|1\rangle &= i\int d\mu_g(x')\,\langle 2\Big|\delta J_1(x')\phi(x)\phi(x') + \delta J_2(x')\phi(x')\phi(x)\Big|1\rangle\\ &= i\int d\mu_g(x')\delta J(x')\,\langle 2\Big|\mathcal{T}\Big(\phi(x)\phi(x')\Big)|1\rangle\,,\end{aligned} \tag{4.149}$$

where the time ordered product is defined with respect to the (arbitrary) Cauchy surface $\Sigma \ni x$. More generally, one must define a (arbitrary) foliation of spacetime between Σ_1 and Σ_2 to concretely define time-ordered products involving any two events. However, note that, since $[\phi(x), \phi(x')]$ vanishes for spacelike separated events, our result will be *foliation-independent*. Finally, we can rewrite (4.149) in terms of functional derivatives:

$$\frac{\delta^2 Z[J]}{\delta J(x_1)\delta J(x_2)} = i^2 \left\langle 2\middle|\mathcal{T}\Big(\phi(x_1)\phi(x_2)\Big)\middle|1\right\rangle . \tag{4.150}$$

And, by induction, it is not hard to generalize to variations of any order:

$$\frac{\delta^n Z[J]}{\delta J(x_1)\dots\delta J(x_n)} = i^n \left\langle 2\middle|\mathcal{T}\Big(\phi(x_1)\dots\phi(x_n)\Big)\middle|1\right\rangle . \tag{4.151}$$

Now, we are in position to demonstrate the equivalence between the Schwinger Action Principle and path integrals. To do so, it is convenient to introduce a "functional index" notation, so that operations with continuous indexes take a similar form to those with discrete ones; this serves both to compactify our expressions and to promptly recognize operator invariants in the continuum as matrix invariants (such as determinants and traces). In this notation, we write spacetime variables as indices, like $\phi(x) = \phi^i$ or $J(x) = J_i$, integrals as implicit sums, such as:

$$\int d\mu_g(x) J(x)\phi(x) = J_i\phi^i, \tag{4.152}$$

and functional derivatives compactly with indices following commas, such as:

$$\frac{\delta S}{\delta\phi(x)} = \frac{\delta S}{\delta\phi^i} = S_{,i} \tag{4.153}$$

(where the functional variable in respect to which it is being derived is left implicit and to be understood from context, just like when one write partial derivatives compactly as $\frac{\partial}{\partial x^i} = \partial_i$).

Thus, we could rewrite (4.151) as:

$$Z^{,j_1\dots j_n} = i^n \langle 2|\mathcal{T}(\phi^{j_1}\dots\phi^{j_n})|1\rangle . \tag{4.154}$$

Our goal here will be to write the transition amplitudes as the functional integral of a kernel of a functional differential equation, just like the path integral is a functional integral of one fundamental kernel (given by the imaginary exponential of the action).

To do so, we write Z as a Taylor series in J around $J = 0$. In ordinary and in compact notations, it reads:

$$\begin{aligned} Z[J] &= \sum_{n=0}^{\infty} \frac{1}{n!} \int \Big(\prod_{j=1}^{n} d\mu_g(x_j) \Big) J(x_1) \dots J(x_n) \frac{\delta^n Z[J]}{\delta J(x_1) \dots \delta J(x_n)} \Big|_{J=0} \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} J_{i_1} \dots J_{i_n} Z^{,i_1 \dots i_n}[J = 0]. \end{aligned} \tag{4.155}$$

Then, using (4.151):

$$\begin{aligned} Z[J] &= \sum_{n=0}^{\infty} \Big(\frac{i^n}{n!} \Big) J_{i_1} \dots J_{i_n} \ \langle 2 | \mathcal{T}(\phi^{j_1} \dots \phi^{j_n}) | 1 \rangle \\ &= \langle 2 \Big| \ \mathcal{T} \Big(\sum_{n=0}^{\infty} \Big(\frac{i^n}{n!} \Big) J_{i_1} \dots J_{i_n} \phi^{j_1} \dots \phi^{j_n} \Big) \Big| 1 \rangle \\ &= \langle 2 \Big| \mathcal{T}(e^{iJ_i \phi^i}) \Big| 1 \rangle \, . \end{aligned} \tag{4.156}$$

Similarly, we can Taylor-expand the action as a function of ϕ around $\phi = 0$:

$$S[\phi] = \sum_{n=0}^{\infty} \frac{1}{n!} S_{,i_1 \dots i_n} \phi^{i_1} \dots \phi^{i_n}, \tag{4.157}$$

as well as its first derivative:

$$S[\phi]_{,i} = \sum_{n=0}^{\infty} \frac{1}{n!} S_{,i\, i_1 \dots i_n} \phi^{i_1} \dots \phi^{i_n}. \tag{4.158}$$

This allows us to write the expression:

$$\begin{aligned} S_{,i}[\phi] e^{iJ_j \phi^j} &= \sum_{n=0}^{\infty} \frac{1}{n!} S_{,i\, i_1 \dots i_n} \phi^{i_1} \dots \phi^{i_n} e^{iJ_j \phi^j} \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} S_{,i\, i_1 \dots i_n} \Big[i^{-n} \frac{\delta^n}{\delta J_{i_1} \dots \delta J_{i_n}} e^{iJ_i \phi^i} \Big] \\ &\equiv S_{,i} \Big[i^{-1} \frac{\delta}{\delta J} \Big] e^{iJ_j \phi^j}, \end{aligned} \tag{4.159}$$

where we have defined the functional differential operator $S[i^{-1} \frac{\delta}{\delta J}]$ in the last line. Then, applying it to our transition amplitudes $Z[J] = \langle 2 | 1 \rangle \, [J]$, we obtain:

$$\begin{aligned} S_{,i} \Big[i^{-1} \frac{\delta}{\delta J} \Big] Z[J] &= \langle 2 \Big| S_{,i} \Big[i^{-1} \frac{\delta}{\delta J} \Big] \mathcal{T}(e^{iJ_j \phi^j}) \Big| 1 \rangle \\ &= \langle 2 \Big| S_{,i}[\phi] \mathcal{T}(e^{iJ_j \phi^j}) \Big| 1 \rangle \\ &= -J_i Z[J], \end{aligned} \tag{4.160}$$

where we have used (4.137) in the last equality.

Then, we want to solve equation (4.160) for the transition amplitude Z through a functional integral kernel $F[\phi]$. Of course, a specific solution should not only depend on the differential equation, but also on a given set of boundary conditions. Rigorously, it is far from trivial to properly define functional integrals, as well as providing appropriate boundary conditions. In what follows, however, we shall ignore these subtleties and quite pragmatically assume that it is possible to define appropriate measures on the functional space to which ϕ belongs, as well as to, impose boundary conditions that make $F[\phi]$ vanish at infinity "for every point in space" (i.e., $F[\phi] \to 0$, as $\phi(x) \to \pm\infty$, for any x in $\mathcal{M}$). Then using some measure $\mu[\phi]$ in space function we write the following "Fourier Transform" for our functional:

$$Z[J] = \int d\mu[\phi] F[\phi] e^{iJ_i\phi^i}, \tag{4.161}$$

in order to solve for the integral kernel F, we enforce equation (4.160):

$$\begin{aligned} 0 &= \int d\mu[\phi]\Big(S_{,i}[\phi] + J_i\Big)F[\phi]e^{iJ_j\phi^j} \\ &= \int d\mu[\phi]\bigg(S_{,i}[\phi]F[\phi]e^{iJ_j\phi^j} - iF[\phi]\frac{\delta}{\delta\phi^i}e^{iJ_j\phi^j}\bigg). \end{aligned} \tag{4.162}$$

Then, integrating the second terms by parts (assuming a boundary condition for which $F[\phi]$ vanishes as ϕ reaches infinity at *any* event x)[X] , we obtain:

$$0 = \int d\mu[\phi]\bigg(S_{,i}[\phi]F[\phi] + iF_{,i}[\phi]\bigg)e^{iJ_j\phi^j}. \tag{4.163}$$

Since this equation must be valid for any event x (for any index i in our compact notation), and for any source J, it implies:

$$S_{,i}[\phi]F[\phi] + iF_{,i}[\phi] = 0 \qquad \Rightarrow \qquad F[\phi] = Ce^{iS[\phi]}, \tag{4.164}$$

being C a normalization constant.

Thus, we finally obtain transition amplitudes in the form of path integrals:

[X] In the light of the result (4.164) below, one particularly convenient way to enforce this boundary condition, which is intimately associated with the Feynman propagator G_F is to add a small negative imaginary contribution to the mass, $m^2 \to m^2 - i\epsilon$ so that S gets an infinite imaginary contribution as $\phi \to \infty$.

$$Z[J] \equiv \langle 2|1\rangle\,[J] = C\int d\mu[\phi]\,e^{i(S[\phi]+J_j\phi^j)}. \tag{4.165}$$

This recovers Feynman's formulation for our modified action (4.136). Particularly, taking $J = 0$, we recover Feynman amplitudes for our original action.

From the relation (4.165), we can immediately find an expression for the effective action as a function of the source J:

$$e^{iW[J]} = Z[J] = \langle 2|1\rangle\,[J], \tag{4.166}$$

which will yield variations with respect to J in the form:

$$\frac{\delta Z}{\delta J(x)} = ie^{iW}\frac{\delta W}{\delta J} \quad \Rightarrow \quad \frac{\delta W}{\delta J(x)} = \frac{\langle 2|\phi(x)|1\rangle}{\langle 2|1\rangle} \equiv \langle\phi(x)\rangle\,. \tag{4.167}$$

Note that here we are using the brackets $\langle\ldots\rangle$ to denote the normalized transition amplitudes between nonorthogonal initial and final states. These will only coincide with ordinary expected values in a fixed state when $|2\rangle = |1\rangle$.

Particularly, we will be interested in the case of a free scalar field, whose action will be bilinear in field operators (and its spacetime derivatives), so that its Taylor expansion on field variables is simply:

$$S_J[\phi] = \tfrac{1}{2}S_{,ij}\phi^i\phi^j + J_i\phi^i. \tag{4.168}$$

This will be a particularly tractable case, because path integrals can then be evaluated as the product of (an infinite number of) Gaussian integrals. Note that we can write an expression for the effective action in the form:

$$e^{iW[J]} = \int d\mu[\phi]e^{i\frac{1}{2}S_{,ij}\phi^i\phi^j + J_i\phi^i}. \tag{4.169}$$

Before we can evaluate this functional integral, let us analyze a simpler analogue of it in a finite-dimensional space. If A is a symmetric operator acting on $\mathbb{R}^n$ – such that its components can be written as $n\times n$ matrices, with components $A_{ij} = A_{ji}$ – we define the following integral:

$$I(A) = \int d^nx\,e^{\frac{i}{2}(x,Ax)} = \int d^nx\,e^{\frac{i}{2}A_{ij}x^ix^j}. \tag{4.170}$$

Since A is symmetric, it can be diagonalized by an orthogonal matrix M: $L = M^T AM$, where $L_{ij} = L_j \delta_{ij}$. Then, we can use M to perform a variable transformation in our space $x \to y = M^T x$, with unit Jacobian, such that our integral reads:

$$I(A) = \int d^n y \prod_{j=1}^{n} e^{\frac{i}{2}L_j (y^j)^2} = \prod_{j=1}^{n}\left(\int_{-\infty}^{+\infty} dy^j \; e^{\frac{i}{2}L_j (y^j)2}\right) = \prod_{j=1}^{n}\left(\frac{2\pi i}{L_j}\right)^{\frac{1}{2}}. \tag{4.171}$$

Well, but this is just a product of the inverse eigenvalues of L, which are the same as the eigenvalues of A; we have that $\prod_j L_j = \det(L) = \det(A)$. Thus, our integral reads:

$$I(A) = (2\pi i)^{\frac{n}{2}} (\det A)^{-\frac{1}{2}} = (2\pi i)^{\frac{n}{2}} \left(\det A^{-1}\right)^{\frac{1}{2}} \tag{4.172}$$

(where we have written the last equality in terms of the inverse operator A^{-1}, $(A^{-1})^{ij} A_{jk} = \delta^i{}_k$).

Then, if we define a new measure μ on $\mathbb{R}^n$, by absorbing constant prefactor $(2\pi i)^{1/2}$:

$$d\mu(x) = \prod_i \frac{dx^i}{(2\pi i)^{1/2}}, \tag{4.173}$$

we obtain simply:

$$\int d\mu(x) e^{iA_{ij}x^i x^j} = (\det A)^{-\frac{1}{2}} = \left(\det A^{-1}\right)^{\frac{1}{2}}. \tag{4.174}$$

Then, back to our infinite-dimensional case, we can take $J = 0$ and perform an analogous Gaussian integral, yielding:

$$e^{iW[0]} = \int d\mu[\phi] e^{i\frac{1}{2}S_{,ij}\phi^i\phi^j} = \det(S_{,ij})^{-\frac{1}{2}}. \tag{4.175}$$

Note that, in defining a suitably normalized (4.173), we avoided a divergent constant prefactor in our infinite-dimensional expression. However, generally, any constant multiplicative factor will yield only a constant additive factor to W, which will make no contribution to its variations.

Now, we can also write this expression in terms of an inverse operator G^{ij}, obeying $G^{ij} S_{,jk} = \delta^i_k$. Such inverse operators are given exactly by the Green functions to our field equations. In our functional case, however, the specific Green function to be used will also depend on a choice of boundary conditions. We shall not dive in the technical details here about these, but, for our particular choice of boundary condition in our path integrals, the appropriate kernel will be given by the Feynman propagator $-G_F$ (the reversed sign is due to equation (2.133)), which obeys:

$$\int d\mu_g(x') \left[(\partial_t^2 + H_{\mathbf{x}}^2)\delta(x,x')\right] G_F(x',x'') = -\delta(x,x'') \tag{4.176}$$

(where we take note that G_F appeared in section 2.5 as the kernel of particular limit of the modified field equations (2.132)). In condensed notation, this reads:

$$S_{,ij}(G_F)^{jk} = -\delta_i^{\ k} \qquad \Leftrightarrow \qquad S_{,ij}(-G_F)^{jk} = \delta_i^{\ k}. \tag{4.177}$$

Then, we can write the effective action $W \equiv W[J=0]$ as:

$$W = -\frac{i}{2}\ln(\det(-G_F)) = -\frac{i}{2}\,\mathrm{Tr}[\ln(-G_F)], \tag{4.178}$$

where we evaluate the trace of an operator $K(x,x')$ in the continuum as:

$$\mathrm{Tr}\,K = \int d\mu_g(x) K(x,x). \tag{4.179}$$

To obtain an operationally useful representation of G_F, we use the following integral representation for a regularized inverse operator:

$$K^{-1} = \lim_{\epsilon\to 0^+}(K - i\epsilon)^{-1} = \lim_{\epsilon\to 0^+}\left\{-i\int_0^\infty ds\, e^{-i(K-i\epsilon)s}\right\}. \tag{4.180}$$

This identity allows one to perform a spectral integral in G_F, so that it can be cast in a the form due DeWitt[Y]:

$$G_F^{DS}(x,x') = -i(4\pi)^{-\frac{n}{2}}\Delta^{\frac{1}{2}}(x,x')\int_0^\infty ds\,(is)^{-\frac{n}{2}} e^{-im^2 s + \frac{\sigma}{2is}} F(x,x';is), \tag{4.181}$$

where n is the dimension of spacetime, $\sigma(x,x')/2$ is the proper geodesic distance between x and x' and $\Delta(x,x')$ is the so-called Van Vleck determinant:

[Y] For more details on the derivation of this expression, see section 3.6 of citebirrell and references therein.

$$\Delta(x,x') = -\det[\partial_\mu \partial'_\nu \sigma(x,x')][g(x)g(x')]^{-\frac{1}{2}}. \tag{4.182}$$

The convenient thing about this expression is that $F(x,x';is)$ can be written in the form of an asymptotic expansion:

$$F(x,x';is) = \sum_{j=0}^{\infty} a_j(x,x')(is)^j, \tag{4.183}$$

where the coefficients $a_j(x,x')$ depend only on geometrical quantities evaluated at the events x and x'. In practice, they are quite complicated to derive in curved spaces, as one must parallel transport various geometric tensors along the geodesics joining x and x'. However, when we use this propagator to compute expectation values of *local* observables, taking $x \to x'$ (going from a well-defined distribution to an ill-defined divergent object) as in the trace (4.178), they will take a considerably simpler form, depending only on local geometric tensors at x. We display the results for the first 3 of them (see eqs. (6.46)-(6.48) of citebirrell):

$$a_0(x) = 1, \tag{4.184a}$$

$$a_1(x) = \big(\xi - 1/6)R(x), \tag{4.184b}$$

$$a_2(x) = \frac{1}{180} R_{\alpha\beta\mu\nu}R^{\alpha\beta\mu\nu} - \frac{1}{180} R_{\mu\nu}R^{\mu\nu} + \tfrac{1}{2}(\xi - 1/6)R^2 + \tfrac{1}{6}(1 - 1/5)\Box R. \tag{4.184c}$$

Then, to cast the effective action in terms of this asymptotic expansion, we note that the logarithm of an operator can similarly be written as:

$$\ln K = \int_0^\infty \frac{e^{iKs}}{is} ids \tag{4.185}$$

(where we ignore an infinite additive constant arising in the lower bound).

Now, identifying G_F with $-K^{-1}$ in the above expressions, we can finally substitute (4.181) in (4.178) to obtain:

$$W = \frac{i}{2} \int d\mu_g(x) \lim_{x' \to x} \left\{ \int\limits_0^\infty ds \, \frac{e^{-i\left(m^2 s + \sigma/(2s)\right)}}{s^3} \left[\sum_{j=0}^{\infty} a_j(x,x')(is)^j \right] \right\}. \tag{4.186}$$

We can then write this expression in terms of an integral of an effective Lagrangian:

$$W = \int d\mu_g(x)\mathscr{L}_{eff}(x), \tag{4.187}$$

where we define

$$\mathscr{L}_{eff}(x) = -\lim_{x'\to x}\left\{\frac{\Delta^{\frac{1}{2}}(x,x')}{2(4\pi)^{\frac{n}{2}}}\int\limits_0^\infty ds\,\frac{e^{-i\left(m^2s+\sigma/(2s)\right)}}{s^{\frac{n}{2}+1}}\left[\sum_{j=0}^{\infty}a_j(x,x')(is)^j\right]\right\}. \tag{4.188}$$

From these expressions, it is possible to verify that there will be two types of divergences in W. The first one will be associated with taking the integral (4.187) in an infinite spacetime volume. This one is relatively easy to manage, as we can still derive meaningful local expressions for its integrand. The second type of divergence are those that appear directly in the effective Lagrangian $\mathscr{L}_{eff}$. These appear when we take the limit $x \to x'$ and will be much more intricate to handle, requiring appropriate procedures of renormalization; in this limit, the damping factor $\sigma(x,x')/(2s)$ will vanish in the integrand (in fact it vanishes in the entire light cone), making the integral divergent in its lower limit.

From our asymptotic expansion, however, we see that, in $n = 4$ spacetime dimensions, it will be only the first 3 terms in the integral (4.188) that will yield divergent contributions as $s \to 0$, the rest of them being regular. We write this divergent contribution as:

$$\mathscr{L}_{div}(x) = -\lim_{x'\to x}\left\{\frac{\Delta^{\frac{1}{2}}(x,x')}{32\pi^2}\int\limits_0^\infty ds\,\frac{e^{-i\left(m^2s+\sigma/(2s)\right)}}{s^3}\left[a_0(x,x')+a_1(x,x')(is)+a_2(x,x')(is)^2\right]\right\}. \tag{4.189}$$

Now, although this is an effective Lagrangian associated with the matter fields, the coefficients a_0, a_1, a_2 only depend on local geometric tensors as $x \to x'$ (see eqs. (4.184)). This fact will allow us to *absorb the divergent terms in the purely geometrical, gravitational action* $\mathscr{L}_G$[Z] , whose "bare" (unrenormalized) form reads:

$$\mathscr{L}_G = \frac{R - 2\Lambda_B}{16\pi G_B}, \tag{4.190}$$

[Z] Actually, due to the quadratic terms that appear in a_2 (eq (4.184c)), one cannot absorb all divergencies in Λ_B and G_B; it is also necessary to consider additional parameters following quadratic terms in curvature. In practice this could yield quantum corrections to GR. We note, however that it is the value of the *renormalized* parameters that should have physical meaning and these should be ultimately determined by comparison with experiment. For this 'corrected' action to be valid in the limits where GR is well tested, though, these quadratic coefficients would have to be relatively small (and, in principle, there is no reason why they could not be zero).

by defining new (renormalized) parameters Λ, G which will be conceived as a correction of Λ_B, G_B by the addition of (infinite) contributions from the divergent terms arisinf in $\mathscr{L}_{div}$.

By doing that, one may define the renormalized matter action as the finite remainder:

$$\mathscr{L}_{ren} \equiv \mathscr{L}_{eff} - \mathscr{L}_{div} \quad = - \lim_{x' \to x} \left\{ \frac{\Delta^{\frac{1}{2}}(x,x')}{32\pi^2} \int_0^\infty ds \, \frac{e^{-i\left(m^2 s + \sigma/(2s)\right)}}{s^3} \left[\sum_{j=3}^{\infty} a_j(x,x')(is)^j \right] \right\}. \tag{4.191}$$

Of course, as we can anticipate from the previous sections, actually carrying the required regularizations and subtractions in $\mathscr{L}_{eff}$ takes very cumbersome and tortuous calculations. Unfortunately, it will remain out the scope of this dissertation to actually derive some of them explicitly and illustrate these analytical procedures of renormalization involving the effective action. For those, we refer the reader to the very thorough section 6.2 of citebirrell, where the methods of dimensional regularization, zeta function regularization and point-split regularization are explicitly derived, and the renormalization of the geometric parameters is thoroughly discussed.

5 STANDARD AND INFLATIONARY COSMOLOGY

In this chapter, we finally pay closer attention to the subject of cosmology, and dwell in some of the ways in which the theoretical framework developed in the preceding chapters may help to elucidate some of the most pressing questions that we have about our own universe.

In section 5.1, we discuss at considerable length the foundations and some paradigmatic results of standard cosmology: we start by thoroughly constructing FLRW spaces and some of its relevant cosmological observables, then, making use of these constructions, we give an overview of how they culminate in the standard cosmological model – the ΛCDM model –, and, finally, we show some of the fundamental issues in it, which motivate the community in the field to posit a primordial inflationary period.

In section 5.2 we qualitatively discuss the bases of field theory which allow for a dynamical description of an inflationary scenario and briefly comment on the related subject of spontaneous symmetry breaking, within the scope of a few simple models.

Finally, in section 5.3, we discuss the bases and some of the developments of inflationary cosmology, within the particular scenario of chaotic inflation. Throughout this section, we consider a simplified model of an interaction scalar field to perform a few concrete computations and draw estimates for some quantities and potential observational predictions of inflation. We begin this analysis with a more thorough discussion of initial conditions, arguing that chaotic inflation should provide a reasonable framework for this matter. We follow by showing how an interacting field ϕ may produce a finite quasiexponential inflation phase as ϕ slowly decays from its unstable vacuum towards a stable one, and, in the sequence, we quantize its linearized perturbations near its slowly varying equilibrium value and show how the spectrum of this perturbation for very long wavelengths may give rise to a (nearly) scale-invariant spectrum in the CMB. Finally we comment briefly on the evolution of the universe after inflation, and how it can take the form of the hot, radiation-dominated universe that we observe (or draw well-verified predictions from) at later times.

5.1 Standard Cosmology: The $\Lambda\mathrm{CDM}$ Model

The origin and development of the universe is something that has raised many questions and speculations throughout the entire history of human civilization. However, our capacity to more closely observe the skyes, as well as our knowledge from laws of nature to systematically analyze our observations has never been nearly as powerful as it has become in the last 100 years. In this section, we shall explore some of the major developments in the field of cosmology that have arisen based on the theory of general relativity, and make for the picture of what is now known as the standard cosmological model.

5.1.1 The Cosmological Principle and FLRW metrics

Attempts to apply the General Theory of Relativity to obtain a meaningful description of (some average properties of) the large scale universe date back to the very first years of General Relativity itself. Early attempts were strongly marked by constraints of simplicity and philosophical considerations, such as that we should not occupy a distinguished position in universe, or live in a distinguished time in its history. This led to foundind hypotheses that the universe was (on average) homogeneous and isotropic, and even that it was eternal. Einstein himself first introduced a Cosmological Constant in his equations in 1917 (32) to allow for a universe that was spacetime homogeneous (*i.e.*, both spatially homogeneous and eternal). A few years later, in 1922, Friedmann arrived at the first cosmological solutions for Einsteins equations which contemplated the possibilities of an expanding or contracting universe (33); Lemaître arrived at the same solutions indepently in 1927 (34). The possibility of an expanding universe, in opposition to a static one, came to be strongly favored after Hubble's observations in 1929 (31) that distant galaxies seemed to be moving apart from us, with receding velocities roughly proportional to their distance. Later, in the 1930's Robertson and Walker rigorously demonstrated that these solutions were indeed unique (up to topological identifications) for a spatially homogenous and isotropic spacetime (35, 36). For all these contributions, these solutions are collectively known as Friedmann-Lemaître-Robertson-Walker (FLRW) spaces.

Let us now detach a little from the historical details, and go through a simple (physically motivated) mathematical construction of the FLRW spaces, starting from the basic assumptions of spatial homogeneity and isotropy. The physical motivation for this simplifying assumption lies in the so-called (modern) cosmological principle, which states that we should not occupy a special position in the universe, or, more generally that there are no distinguished positions in it. We can roughly summarize this as follows: "for each instant in time, every point in space should look the same". Similarly, there should be no distinguished directions in space, that is: "in every point in space at any instant of time, every spatial direction should look the same". Of course, such considerations do not apply at *any* scales in our universe. It is evident that from subatomic and astrophysical scales the universe is highly inhomogeneous and anysotropic, particularly since the gravitational collapse of matter creates many types of structures at considerably large scales. Nevertheless, these hypotheses turn out to apply very well on *very large, cosmological scales*, and increasingly so as we go backwards in time, as matter becomes less and less gravitationally clumped. Moreover, besides the simplicity, physical appeal and applicability of those hypotheses, there is the further advantage that they result in a cosmological model with very few parameters to be adjusted, as it will be tightly constrained by symmetries. Thus, it is truly remarkable that the ΛCDM can explain so accurately our most precise cosmological observations up to this date.

Now that we have informally stated intuitive notions of spacial homogeneity and isotropy, let us formulate these in a mathematically precise manner, through geometrical restrictions in our spacetime.

We start with the notion of spatial homogeneity: a spacetime $(\mathcal{M}, g_{ab})$ is said to be spatially homogeneous if there is a 1-parameter-family of spacelike surfaces Σ_t foliating $\mathcal{M}$ such that, given a time instant t and any 2 points $p, q \in \Sigma_t$, there is an isometry $\mathcal{I}$ ($\mathcal{I} : (\mathcal{M}, g_{ab}) \to (\mathcal{M}, g_{ab})$) that takes p into q, $\mathcal{I}(p) = q$ (See Figure 40).

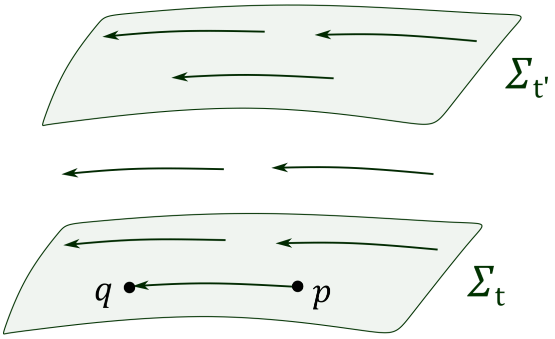

Figure 40 – Illustration of a translation isometry in a spatially homogeneous spacetime, which takes the point $p \in \Sigma_t$ and maps it into $q \in \Sigma_t$. This picture also tries to convey the fact that a translation isometry acts *in the entire spacetime*, but mapping each homogeneous surface Σ_t in itself.
Source: By the author.

As to spatial isotropy, we must first emphasize that, for a general spatially homogeneous spacetime, there can be at most one observer in each event p that 'sees space around him as isotropic'; correspondingly, there will be at most one foliation Σ_t of $\mathcal{M}$ which will be everywhere spatially isotropic (and homogeneous). This "isotropic observer" will be the one whose worldline is orthogonal to the hypersurface Σ_t at p, such that his 'spatial directions' all lie parallel to Σ_t. Let then u^a be the tangent vector to his worldline at p and let s_1^a and s_2^a be any two normalized purely spatial vectors to him (*i.e.*, tangent to Σ_t at p, such that $u_a s_i^a = 0$); a spacetime will be said spatially isotropic *at a point* p if there is an isometry $\mathcal{I}$ preserving p and u^a and rotating two arbitrary normalized spatial vectors s_1^a and s_2^a into one another, $\mathcal{I}^*(s_1^a) \to s_2^a$ [A] (see Figure 41).

[A] $\mathcal{I}^*$ denotes the *pushforward* map induced by $\mathcal{I}$ in vectors tangent to M. For more details on diffeomorphisms between manifolds and their induced maps on tangent tensor fields, see appendix B.

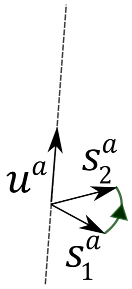

Figure 41 – Illustration of the action of the pushforward map I^*, associated with a spatial rotation isometry. Here I^* rotates the normalized spatial vector s_1^a in another such vector s_2^a.
Source: By the author.

Now the imposition that we wish to make (on the basis of the cosmological principle) is that $\mathcal{M}$ is spatially isotropic *at all of its points*. As we will demonstrate next, this will be a particularly restrictive condition for spacetime geometry. Let $h_{ab}(t)$ be the (positive-definite) metric induced in Σ_t by g_{ab}[B]; we may use the covariant derivative D_a associated to it (i.e. $D_a h_{bc} = 0$) to construct a spatial curvature tensor on Σ_t: ${K_{abc}}^d$, and then raise its 3rd index with the metric h_{ab}: ${K_{ab}}^{cd} = h^{ce}{K_{abe}}^d$. Due to the antissimetry properties of Riemann curvature on the first and second pair of indices ($K_{abcd} = -K_{bacd} = K_{badc}$), ${K_{ab}}^{cd}$ can be thought of as a map L between 2-forms (antisymmetric rank (0,2) tensors) in this subspace:

$$\begin{aligned} {K_{ab}}^{cd} \to L :\ & W \to W \\ & \omega_{ab} \to {K_{ab}}^{cd}\omega_{cd}\ (\equiv L\omega). \end{aligned}$$

Further, h^{ab} may be used to define an inner-product H between 2-forms.

$$\begin{aligned} H :\ & W \times W \to \mathbb{R} \\ & (\omega_{ab}, \mu_{cd}) \to h^{ac}h^{bd}\omega_{ab}\mu_{cd}\ (\equiv \langle \omega, \mu \rangle), \end{aligned}$$

and it is easy to see that L will be a symmetrical (self-adjoint) map with respect to H: $\langle \omega, L\mu \rangle = \langle L\omega, \mu \rangle$. Thus, there will be in W a basis of eigenvectors ("eigen-2-forms") of L. The restriction of spatial isotropy will then imply that the eigenvalues of L *must all be the same*, otherwise, one could use this purely geometrical prescription to build distinguished 2-forms, and thus distinguished planes and directions in Σ_t (more concretely, we can interpret that different eigenvalues would result in planes with different curvatures tangent to Σ_t). Thus, L must act as a multiple of the identity operator in W (and annihilate all symmetrical rank (0,2) tensors):

[B] In the entire spacetime, h_{ab} can be seen as a projector (with an inverse sign for our $(+,-,-,-)$ choice of signature) on the tangent spaces parallel to each Σ_t: $h_{ab} = -(g_{ab} - u_a u_b)$.

$$L = \kappa\, \mathbb{1}_W \quad \Leftrightarrow K_{ab}{}^{cd} = \kappa\, \delta^c_{[a}\delta^d_{b]} \quad \Leftrightarrow K_{abcd} = \kappa\, h_{c[a}h_{b]d}. \tag{5.1}$$

Further, spatial homogeneity will imply that κ must be a constant throughout each Σ_t. Curiously, this homogeneity actually turns out to be a necessary consequence of isotropy *at all points*, which can be demonstrated by the fact that the curvature tensor K_{abcd} must obey a Bianchi identity:

$$0 = D_{[e}K_{ab]cd} = (D_{[e}\kappa)h_{|c|a}h_{b]d}, \tag{5.2}$$

and, for a manifold Σ of dimension 3 (or larger), the rightmost side of this equation will be null if, and only if, $D_e\kappa = 0$ (*i.e.*, if κ is a constant in each Σ_t).

Now, any two isotropic spaces of the same constant curvature κ will be locally isometric. The problem of finding instantaneous possible solutions to Einstein equations with such symmetries then reduces to that of classifying all possible 3-dimensional geometries with constant isotropic curvature. As Robertson and Walker first demonstrated in the 1930's (35,36), there are only 3-possibilities if we assume a usual, simply-connected topology, corresponding to the spatial metrics:

$$d\Sigma^2 = \begin{cases} d\psi^2 + \sin^2(\psi)[d\theta^2 + \sin^2\theta d\phi^2] & (\Sigma = \mathbb{S}^3), \quad \kappa > 0 & \text{(5.3a)} \\ d\psi^2 + \psi^2[d\theta^2 + \sin^2\theta d\phi^2] & (\Sigma = \mathbb{R}^3), \quad \kappa = 0 \quad , & \text{(5.3b)} \\ d\psi^2 + \sinh^2(\psi)[d\theta^2 + \sin^2\theta d\phi^2] & (\Sigma = \mathbb{H}^3), \quad \kappa < 0 & \text{(5.3c)} \end{cases}$$

where we have written the line element in spherical coordinates with a proper-distance radial coordinate ψ.

Along with the orthogonal contribution in the isotropic timelike directions, and accounting for a possible time-dependence in the space metric, this gives us the total spacetime metric:

$$ds^2 = dt^2 - a^2(t)d\Sigma^2 = dt^2 - d\Sigma_t^2. \tag{5.4}$$

Often, it is more convenient to write all three possibilities for the spatial metric compactly in terms of the areal radial coordinate r, in terms of the normalized curvature $k = a^2\kappa$, so that ds^2 reads:

$$ds^2 = dt^2 - a^2(t)\left[\frac{dr^2}{1-kr^2} + r^2\Big(d\theta^2 \sin^2(\theta)d\phi^2\Big)\right], \qquad \begin{cases} k = +1 & \Rightarrow \Sigma = \mathbb{S}^3 \\ k = 0 & \Rightarrow \Sigma = \mathbb{R}^3 \\ k = -1 & \Rightarrow \Sigma = \mathbb{H}^3 \end{cases} . \tag{5.5}$$

Note that, in the spatially flat case, the instantaneous value of $a(t)$ at any particular time t_0 does not have direct physical meaning, and it can always be redefined in a change of spatial coordinates (in (5.5), this would be $r \to a(t_0)r$); in this case, the only physically meaningful quantity is its relative time variation $H = \dot{a}/a$. However, in the spatially curved cases (either spheric or hyperbolic) $a(t)$ *directly* provides a physical length scale in the universe, namely, the one associated with the inverse spatial curvature in Σ_t (this can be quite intuitively associated with the 'radius' of the universe for the spherical case, but, although the hyporbolic case has a noncompact spatial section, *both* curved geometries have intrinsic geometrical observables that reflect their curvature scales. This will become more evident in the Friedmann equations).

Note further that all considerations so far have not specifically assumed GR (except insofar as we are assuming spacetime to have a specific structure of a 4D pseudo-Riemannian manifold with a Lorentzian metric): we have not yet imposed Einstein's equations. Such features will then be common to any gravity theories that share this basic structure, as long as we constrain the analysis with the very strict symmetry hypothesis of perfect spatial homogeneity and isotropy.

Having analyzed some basic geometrical features of a spatially homogeneous and isotropic universe, and arrived in the general form of a FLRW metric, we would now like to substitute the general metric (5.5) in Einstein's equations to derive predictions about the dynamical evolution of our universe – *i.e.* to obtain an explicit form of $a(t)$ given a distribution of matter and energy. Therefore, to do so, let us begin by making a few considerations about matter and energy content in a homogeneous and isotropic universe.

We begin by noting that the most general matter/energy distribution which is fully consistent with our hypotheses of homogeneity and isotropy will take the form of a perfect fluid (that is, a fluid without viscosity or heat transfer and with null velocity as seen by isotropic observers), less of any terms directly proportional to curvature. One can see why that is by noting that, in that case, the stress tensor for matter can only be built using the metric g_{ab} and the timelike vector field u^a tangent to the isotropic worldlines. Thus, the most general symmetrical rank 2 tensor we can build is of the form[C]:

$$T_{ab} = \alpha u_a u_b + \beta g_{ab}, \tag{5.6}$$

and, using the standard identifications of energy density $\rho = T_{ab}u^a u^b$ and pressure $p = T_{ab}s^a s^b$ (where, again, s^a is *any* normalized spatial vector tangent to the isotropic space

[C] In principle, if one allows terms proportional to curvature, he/she could also add terms proportional to R_{ab}, Rg_{ab} and $Ru_a u_b$ without violating general covariance or the cosmological principle. However, it is highly unusual to consider such forms of *matter* density that relate directly to curvature. Furthermore, *in the right combination*, these terms could be *partly* shoved to the LHS of Einstein's Equations, redefining G.

dections Σ_t), one can easily cast (5.6) in the form:

$$T_{ab} = \rho u_a u_b + p(u_a u_b - g_{ab}), \tag{5.7}$$

which is that of a general perfect fluid. Note that the homogeneity condition further restricts ρ and p to be (at most) functions of time.

In practice, for a cosmological analysis, it will be generally convenient to split T_{ab} in different components to account for different types of matter with different equations of state (that will dictate how p and ρ are related in equilibrium conditions). Such components can be well approximated as noninteracting, for an appreciable part of the history of the universe.

A particularly simple component, which has been dominant for a significant part of the history of our universe, is given by nonrelativistic/cold matter, which is very well modeled as a pressureless fluid $T^{dust}_{ab} = \rho u_a u_b$; this component is very commonly known as 'dust'. Another significant component is the one given by ultrarelativistic energy contributions – most proeminently in the form of electromagnetic radiation, although this would equally apply to any massless particles/fields – whose equation of state in an isotropic distribution is just $p = \rho/3$.

Then, we would like to evaluate Einstein's equation (3.1) (with $\Lambda = 0$) for a FLRW metric (5.3) with a source of the type of a perfect fluid (5.7) with a given equation of state, so that we may solve for $a(t)$, $\rho(t)$ and $p(t)$. In this highly symmetric metric, one can show (with similar arguments as those for the spatial curvature) that the purely spatial portion of the Ricci tensor ${R_a}^b$ must be proportional to the identity operator (${}^{(3)}{\delta_a}^b$) on the subspaces tangent to Σ_t, and that its space-time components should vanish, so that we end up with only two independent components: the time-time (R^t_t) and the (isotropic) space-space (R^i_i) ones. Thus, the independent components of (3.1) read **(in geometrized or Planck units. Maybe insert a planck mass M_p here and unify the notation with section 5.3)**:

$$G_{tt} \equiv R_{tt} - \tfrac{1}{2} g_{tt} R = -8\pi T_{tt} = 8\pi\rho, \tag{5.8a}$$

$$G_{**} \equiv R_{**} - \tfrac{1}{2} g_{**} R = -8\pi T_{**} = 8\pi p, \tag{5.8b}$$

where we have denoted the normalized spatial components of the tensors with an asterisk, $T_{**} = T_{ab} s^a s^b$. In terms of coordinate components, they are simply $T_{**} = (g_{ii})^{-1} T_{ii}$.

Now, to actually evaluate those equations, we must first obtain the Ricci Tensor explicitly in terms of $a(t)$ to put it in the LHS of the equations. These calculations are somewhat lengthy and quite mechanical, so that it is generally useful to obtain them from a symbolic calculator software. Their results for the Ricci components read:

$$R_{tt} = 3\frac{\ddot{a}}{a}, \tag{5.9a}$$

$$R_{**} = (g_{ii})^{-1}R_{ii} = -\left[\frac{\ddot{a}}{a} + 2\left(\frac{\dot{a}}{a}\right)^2 + \frac{2k}{a^2}\right]. \tag{5.9b}$$

From those components, it is easy to compute the Ricci scalar:

$$R = 6\left[\frac{k}{a^2} + \left(\frac{\dot{a}}{a}\right)^2 + \frac{\ddot{a}}{a}\right]. \tag{5.10}$$

Substituting these in (5.8), we finally obtain the famous Friedmann equations:

$$\frac{\dot{a}^2}{a^2} = \frac{8\pi\rho}{3} - \frac{k}{a^2}, \tag{5.11a}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi}{3}(\rho + 3p). \tag{5.11b}$$

Before analyzing in further detail the dynamical consequences of these equations, let us see how the matter and energy content should evolve subject to them when we have a simple equation of state. Multiplying equation (5.11a) by a^2 and taking a time derivative we get:

$$2\dot{a}\ddot{a} = \frac{8\pi}{3}(\dot{\rho}a^2 + 2\rho a\dot{a}) \qquad \Rightarrow \qquad \dot{\rho} + \left(2\rho - \frac{3}{4\pi}\frac{\ddot{a}}{a}\right)\frac{\dot{a}}{a} = 0 \tag{5.12}$$

and, then, using (5.11b):

$$\dot{\rho} + 3(\rho + p)\frac{\dot{a}}{a} = 0. \tag{5.13}$$

Generally, we must supply additional information regarding the relation between p and ρ (*i.e.* an equation of state) so that we may derive the full joint evolution of spacetime and matter. Since for a great part of the history of the universe we may treat it as dominated by a single component, a very simple and widely applicable class of equations of state will be given by a simple proportionality relation of the form $p = w\rho$[D], being w a constant. In this case, we have:

$$\dot{\rho} + \alpha\rho\frac{\dot{a}}{a} = 0, \tag{5.14}$$

[D] In the literature, w itself is commonly called the 'equation of state'.

where we have defined the proportionality constant $\alpha \equiv 3(1+w)$. Now, it is easy to see that (5.14) simply expresses a conservation law in the form:

$$\frac{d}{dt}(\rho a^{\alpha}) = 0 \qquad \Rightarrow \qquad \rho a^{\alpha} = cte. \tag{5.15}$$

Particularly, for a spatially flat universe ($k = 0$), this will yield a very simple power-law solution to eq (5.11a) when $w \neq -1$, $a(t) \propto t^{\lambda}$, where:

$$\lambda = \frac{2}{3(1+w)} \qquad \Leftrightarrow \qquad w = \frac{2-3\lambda}{3\lambda}, \tag{5.16}$$

whereas the $w = -1$ case yields an exponential solution $a(t) \propto e^{Ht}$.

Particularly, for dust and radiation (for which $w = 0$ and $w = 1/3$, respectively), we finde:

$$\begin{cases} \rho_{dust} \propto a^{-3}, \qquad a(t) \propto t^{2/3} & (5.17a) \\ \rho_{rad} \propto a^{-4}, \qquad a(t) \propto t^{1/2} & (5.17b) \end{cases} .$$

In general, one can see that, for any form of matter with positive energy densities $\rho > 0$ and nonnegative pressures $p \geq 0$ (or, equivalently, $w \geq 0$ if we already assume the first equality), we have that ρ will always decay at least as fast as a^{-3}. For $w = 0$, this decay can be thought of as a conserved *total energy* being diluted in a volume that scales as a^3 in the expansion. For $w > 0$, not only is the energy diluted but the positive pressure also *performs work onto the expansion*, causing a corresponding decrease in the total energy.

Two other important contributions in the Friedmann equations, which can play the role either of an *actual* energy component or an *effective* one, are terms associated with spacial curvature and terms proportional to the metric (the latter, as we have seen in chapters 3 and 4, can arise either as a modification in GR by the insertion of a cosmological constant or as term due to vacuum energy[E]); they will have equations of state $w = -1/3$ and $w = -1$. Thus, both these species *receive work from the expansion*, and their (effective) energy densities will behave as:

$$\begin{cases} \rho_k \propto a^{-2} & (5.17c) \\ \rho_\Lambda \propto a^0 = cte & (5.17d) \end{cases} .$$

[E] In fact, this property of Λ, namely, that it can be inserted in either side of Einstein's equation and interpreted either as a geometrical modification of GR or a matter source, was exactly what we used in chapter 4 to take divergent terms in $\langle T_{ab} \rangle$ and absorb them in the renormalization of gravitational constants G and Λ (either incorporating divergencies or finite corrections).

Both of these terms then entail (effective) energy densities that tend to become dominant over ordinary matter/energy forms. In fact, it is precisely the latter form that Dark Energy assumes today, being the dominant energy contribution for roughly the last 4 billion years in the history of our universe, and currently corresponding to about 70% of our total energy density. As of spatial curvature, it seems to be neglegible in any of our cosmological observations, all of which point to a vanishing value of k. We will discuss this term in more detail in the next section.

A further observation is that these ordinary conditions of nonnegative energy density and pressures $\rho, p \geq 0$ imply that *a nonempty universe* ($\rho > 0$) *cannot be static*; particularly, if we look at eq. (5.11b), we see that it must always be *decelerating*, as it is expected from the purely attractive character of gravity observed in subcosmological scales (this last conclusion remains unchanged if we add a nonzero spatial curvature term, as one needs $w < -1/3$ to reverse the sign in (5.11b)). Under these conditions, one concludes that (i) the universe must be dynamic and if it is expanding ($\dot{a} > 0$), it leads to a singularity in a finite time to the past of about $\sim H_0^{-1}$, the well-known *Big Bang* (H_0 being the present value of $H \equiv \dot{a}/a$). To the future, it can either continue to expand indefintely or recollapse, depending on the value of k. We summarize all three scenarios (with no exotic energy components) in Figure 42.

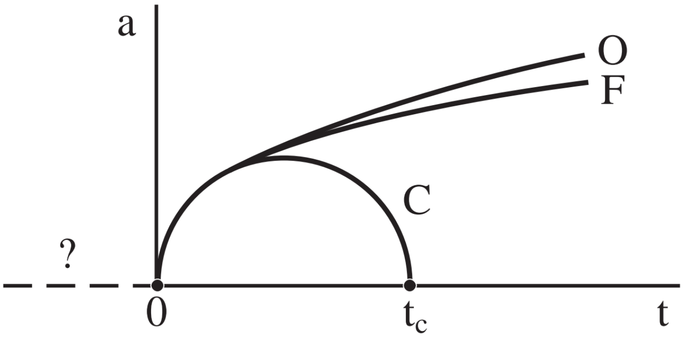

Figure 42 – Scenarios for FLRW universe filled with ordinary matter ($\Lambda = 0$, and likewise for other exotic contributions with $w < -1/3$) for a spherical (Closed, $k=+1$), Flat ($k=0$) or hyperbolic (Open, $k=-1$) universe. In the spherical case, the universe has a finite spatial volume and recollapses in a finite time t_c, whereas in the flat and hyperbolic cases, it has an infinite volume and expands indefinetely.
Source: LINDE (4)

The situation can be considerably different when we consider a term in the form of a cosmological constant (regardless of whether it corresponds to an actual energy term or a modification in Einstein equations). Since it has an equation of state $w = -1$, it can actually lead to an *accelerated* expansion (as it has been occurring in our universe for the last 4 billion years). In fact, if one perfectly balances it with an ordinary contribution in a positive curvature scenario, one could actually find a static solution. This was, in fact, Einstein's original motivation to insert Λ in his equations. Going from (3.1) to (3.2), we correspondingly modify the Friedmann equations to:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{\Lambda + \pi\rho}{3} - \frac{k}{a^2}, \tag{5.18a}$$

$$\frac{\ddot{a}}{a} = \frac{\Lambda}{3} - \frac{4\pi}{3}(\rho + 3p), \tag{5.18b}$$

for which we can obtain a static universe $\dot{a} = 0 = \ddot{a}$ if $\Lambda > 0$ and $k = +1$. It amounts to setting:

$$\Lambda = 4\pi(\rho + 3p), \qquad \text{and} \qquad a = \left[\frac{3k}{\Lambda + 8\pi\rho}\right]^{1/2} = \Big[4\pi(\rho + p)\Big]^{-1/2}. \tag{5.19}$$

This is known as Einstein's Static Universe (ESU). Notice that it is even more symmetric that a generic FLRW universe: it is *spacetime* homogeneous and spatially isotropic (and it is still *locally* symmetric by boosts, but not globally, due its spherical geometry).

To conclude this section, we want to make more explicit the notion of an expanding (or contracting) universe, in terms of the expansion (or contraction) of (instantaneous) scale distances of the isotropic spacelike surfaces associated to the isotropic cosmological frame. Note that, if at a certain time t, the distance between two isotropic observers (*e.g.*, 2 galaxies at rest at the cosmological frame) is $R = ra(t)$ (which we can describe by fixed spherical spatial coordinates $(0, 0, 0)$ and (r, θ, ϕ)), then the rate of variation of that distance will be:

$$\frac{dR}{dt} = r\dot{a} = R\frac{\dot{a}}{a} \equiv RH, \tag{5.20}$$

which is directly proportional to the *intantaneous* geometrical distance R, and the proportionality factor is merely the fractional rate of expansion of the universe, H, which in cosmology is called the *Hubble Parameter.*

As we have briefly stated earlier, Hubble was the first one to make observations that distant galaxies were in fact moving away from us, with a velocity proportional to their distances. Of course, actual measurements will not correspond to instantaneous geometrical distances and velocities, but actually the ones along our past light cone, as light signals emmited from these galaxies take a finite time to reach us, and they will take different times for different distances, corresponding to different values of $H(t)$. Still, for sufficiently close galaxies $R \ll H_0^{-1}$, the measured velocities will approximately obey a simple proportionality relation:

$$\frac{dR}{dt} \simeq H_0 R, \tag{5.21}$$

which was indeed found in Hubble's observations in the late 1920's (31), which played an important role in the realization that our universe is in fact expanding.

5.1.2 Cosmological Parameters, the ΛCDM model and the Hot Big Bang scenario

Although the formulation given so far is very useful to describe and calculate predictions for the evolution of our universe, it mostly refers to quantities that are very difficult (or even impossible) to observe directly. In order to draw a closer connection to quantities that are actually observed, which allow us to constrain our cosmological models and test predictions, we start this subsection by constructing a few observationally-oriented cosmological parameters.

A first remark in what concerns cosmological observations is that we do not have causal access to the entirety of spacetime. Obviously, we do not have access to our causal future, nor to spacelike separated regions (particularly, to any instantaneous geometrical distances in Σ_{t_0}), so that our observations are bounded to probe only the region within our past light cone. In fact, since the striking majority of the information we obtain is transported via electromagnetic radiation, our observation space is virtually restricted to a very narrow window *on* our past light cone (since the time scales of any human observations are extremely small compared to those of cosmological phenomena, the 'temporal thickness' of the set of past light cones comprising an observation – or a series of observations – is neglible for all practical purposes).

As we have mentioned in the previous subsection, radiation propagating in an expanding universe is redshifted. The portion of the redshift effect that is due to cosmic expansion (rather than to any peculiar velocities of the source or the observer relative to the isotropic worldlines) is called the *cosmological redshift*. If we consider a light ray emitted (by an isotropic source) at an event P_1, with wavelenght λ_1, and observed (by an isotropic observer) at an event P_2, with wavelength λ_2, we define this redshift by:

$$z_{21} \equiv \frac{\lambda_2 - \lambda_1}{\lambda_1} = \frac{\omega_1}{\omega_2} - 1. \tag{5.22}$$

One particularly convenient way to compute this redshift is by making use of the translation isometries in this spacetime. Let k^a be the null tangent vector to the propagation of the light ray, we pick a translation Killing field ξ^a that is proportional to its projection in the subspaces tangent to Σ_t, that is (see Figure 43):

$$\xi^a = -h^a{}_b k^b \propto k^a - (u_b k^b) u^a. \tag{5.23}$$

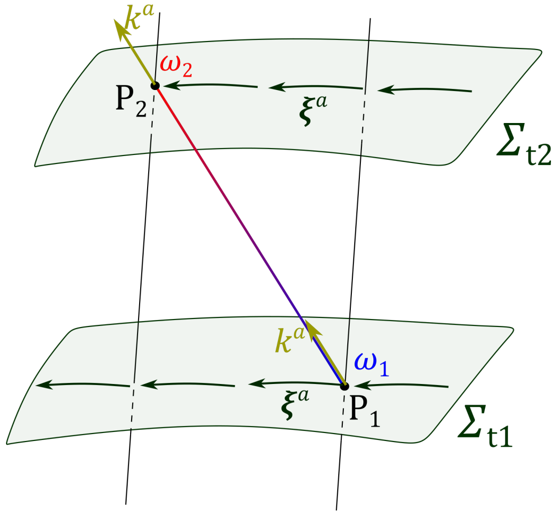

Figure 43 – Depiction of a light-ray emmited at an event P_1, with a frequency ω_1 and absorbed at an event P_2, with a frequency ω_2, as well as the corresponding translation Killing field ξ^a that joins the isotropic worldlines passing through these two events. The cosmological redshift that the light undergoes between those events can be easily computed through the conserved quantity $k_a\xi^a$.
Source: By the author.

This corresponds to the Killing field that generates the translation that joins the wordlines of the source and the observer. Regardless of the field global normalization, it must scale in time as $a(t)$, as it represents a *spacetime isometry*, and thus it must preserve spacetime distances and angles; in particular since it maps points on each simultaneity surface Σ_t to points on *the same* Σ_t, it must take isotropic worldlines into isotropic worldlines, to preserve their arclength (*i.e.*, their proper-time intervals) between *any* two surfaces Σ_{t_1} and Σ_{t_2}. Thus:

$$\frac{|\xi_a\xi^a|^{1/2}|_{P_1}}{|\xi_b\xi^b|^{1/2}|_{P_2}} = \frac{a(t_1)}{a(t_2)}. \tag{5.24}$$

Then, making use of the conserved quantity (along the propagation of the light-ray) $k_a\xi^a$ and using that k^a is a null vector, $k_ak^a = 0$, so that its dispersion relation is simply $\omega = |\mathbf{k}|$ (*i.e.*, $k^au_a = -k^a\xi_a/||\xi||$), we obtain:

$$\omega_1 = k_au_1^a = -k_a\left[\xi^a|\xi_b\xi^b|^{-1/2}\right]_{P_1}, \tag{5.25}$$

$$\omega_2 = k_au_2^a = -k_a\left[\xi^a|\xi_b\xi^b|^{-1/2}\right]_{P_2}. \tag{5.26}$$

Thus, eq (5.24) immediately yields the frequency ratio:

$$\frac{\omega_2}{\omega_1} = \frac{|\xi_a\xi^a|^{1/2}|_{P_1}}{|\xi_b\xi^b|^{1/2}|_{P_2}} = \frac{a(t_1)}{a(t_2)}, \tag{5.27}$$

and the cosmological redshift (5.22) between any 2 events is found simply by:

$$z = \frac{a(t_2)}{a(t_1)}. \tag{5.28}$$

This very simple correspondence between redshift and the scale factors allows one to quite directly trace back the history of expansion of the universe by making a large number of measurements from various sources at different distances (provided one has an independent way to measure distances for distant objects[F]). Furthermore, if the scale factor a is a monotonic function of time (in our case, a monotonically increasing function of time), z is at a one-to-one corresponce with t, making it a directly observable 'time parameter' on our past light cone. Being t_0 the present time and $a_0 \equiv a(t_0)$ our present scale factor, this relation yields:

$$z(t) = \frac{a_0}{a(t)}, \qquad t \leq t_0, \tag{5.29}$$

for which we can find an inverse $t(z)$ by inverting t as a function of a (this is particularly simple in the cases of power-law and exponential expansions).

With this construction, it is observationally more convenient to express other observables and parameters as a function of z, rather than a. Two important geometrical parameters, which appear directly in the Friedmann equations, are the Hubble parameter H and the deceleration parameter q:

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi}{3}\rho - \frac{k}{a^2}, \tag{5.30}$$

$$q \equiv -\frac{\ddot{a}a}{\dot{a}^2} = -\frac{\ddot{a}/a}{H^2}. \tag{5.31}$$

(Historically, q was defined with a negative sign precisely due the expectation that the cosmic expansion should necessarily be *decelerated*, so that one would always have $q > 0$. However, as Dark Energy actually caused our expansion to be *accelerated*, we presently have $q = q_0 < 0$.)

Now, in order to extract meaningful information in terms of cosmological redshifts, one must also be able to measure distances independently. For this reason, we give here a few definitions of cosmic distances throughout spacetime, which turn out operationally or

[F] Indeed there are many ingenious way to define and measure distance at various cosmic scales, based on objects of known luminosity or size (which are respectively known as 'standard candles' and 'standart rules' in the literature), and for which one uses known methods on smaller scales to calibrate measurements on larger ones; this is known as the *cosmic distance ladder* (see *e.g.* (12,13)). Ahead, we shall define a few operationally useful notions of distance.

observationally useful. The first and most obvious definition of distance we could define are instantaneous geometrical distances. For two worldlines with fixed radial distance coordinates 0 and χ, respectively, this will be simply given by:

$$d_G(t) \equiv a(t)\chi. \tag{5.32}$$

We note that it is in terms of this distance that we have written the *geometrical* Hubble Law ((5.20)):

$$\frac{d}{dt}d_G(t) = H(t)d_G(t). \tag{5.33}$$

Although this particular notion of distance is geometrically intuitive, it does not have any direct observational relevance. A more useful notion of distance could be defined through our past light cone, considering the geometrical distance *at the time t of emission of a light ray*, which is subsequently detected in the present, t_0:

$$d_{light}(t) = a(t)\int\limits_{t}^{t_0}\frac{dt'}{a(t')} = \frac{1}{1+z}\int\limits_{0}^{z}\frac{dz'}{H(z')} \tag{5.34}$$

(where we have switched variables $t \to z(t)$ along the past light cone).

If one is able to invert equation (5.34), and obtain $z(d_{light})$, it is possible to use this relation to write an *observational* Hubble Law, in terms of d_{light} and the Hubble *constant*, $H_0 = H(z=0)$ (employing (5.30)). Although an exact inverse is not generally analytically possible, it is quite simple to linearize the equation and obtain a first-order approximation for this law:

$$z = H_0 d_{light} + \mathcal{O}\Big((H_0 d_{light})^2\Big), \tag{5.35}$$

which should be a good approximation well inside the Hubble radius, $d_{light} \ll H_0^{-1}$.

For completeness, we also mention that observationally direct notions of distance can be defined in close relation to the measured light intensity $\mathcal{I}$ of objects of known luminosity L, or the measured angular amplitude $\delta\theta$ of objects of known size δl. Since in Euclidean geometry these quantities relate to the geometrical distance r respectively as:

$$\mathcal{I} = \frac{L}{4\pi r^2}, \qquad \delta\theta = \frac{\delta l}{r}, \tag{5.36}$$

we *define* the cosmic distances (in our expanding, potentially curved) associated to these measurements as:

$$d_L \equiv \sqrt{\frac{L}{4\pi \mathcal{I}}}, \qquad d_A \equiv \frac{\delta l}{\delta \theta}. \tag{5.37}$$

These definitions may seem somewhat awkward in a geometric perspective, but the relevant point is that they can each be *independently* measured in terms of observationally accessible quantities, and then used to compare and test predictions in our cosmological model.

Besides these geometrical parameters, it also proves convenient to define a few parameters for matter, in terms of which we can cast the Friedmann equations in a more observationally convenient form. First, we note that the dimensionless spatial curvature k is not a free parameter, but is rather determined by the energy density. To see this more clearly, we note that if we are in the spatially flat case, $k = 0$, then (5.30) demands that the energy density assumes a very particular value in respect to the expansion rate, called the *critical density*:

$$\rho_c \equiv \left(\frac{3H^2}{8\pi} \right), \tag{5.38}$$

in terms of which we define the relative energy density:

$$\Omega \equiv \rho / \rho_c. \tag{5.39}$$

Then, we immediately have that:

$$\begin{cases} \Omega < 1 \;\Leftrightarrow \rho < \rho_c, & k = -1 \quad \text{(hyperbolic)} \\ \Omega = 1 \;\Leftrightarrow \rho = \rho_c, & k = 0 \quad\;\;\, \text{(flat)} \\ \Omega > 1 \;\Leftrightarrow \rho > \rho_c, & k = +1 \quad \text{(spheric)} \end{cases} . \tag{5.40}$$

As mentioned in the previous subsection, a useful way to decompose the total energy is to consider several components with partial densities ρ_j, such that $\sum_j \rho_j = \rho$, each with a constant equation of state $w_j = p_j/\rho_j$. One may then quite naturally define partial relative densities Ω_j for each component as:

$$\Omega_j \equiv \rho_j / \rho_c, \qquad \sum_j \Omega_j = \Omega, \tag{5.41}$$

as well as an *effective partial density* associated with the curvature term:

$$\Omega_k \equiv -\frac{k}{a^2 H^2} \tag{5.42}$$

(which we have already seen to correspond to an equation of state $w_j = -1/3$).

One can also define an effective equation of state associated to the *total energy density*:

$$w \equiv \frac{p}{\rho} = \frac{\sum_j \rho_j w_j}{\sum_j \rho_j} = \frac{\sum_j \Omega_j w_j}{\Omega}, \tag{5.43}$$

although, clearly, w will not generally be a constant throughout cosmic evolution even if each w_j is (but it will be approximately constant whenever one single component j dominates over all the others; then $w \simeq w_j$).

With these definitions, one can very conveniently rewrite the Friedmann equations (5.11) as:

$$\Omega + \Omega_k = 1, \tag{5.44a}$$

$$q = \frac{\Omega}{2}(1 + 3w). \tag{5.44b}$$

Further we can quite simply express the partial densities in terms of z and $\{w_j\}$. Recall that for noninteracting energy components, we had a series of individual conservation laws:

$$\rho_j a^{3(1+w_j)} = cte. \tag{5.45}$$

In terms of the present partial energy densities ρ_{j0} and the redshift z, these can be solved simply as:

$$\rho_j(z) = \rho_{j0}(1 + z)^{3(1+w_j)}, \tag{5.46}$$

from which we immediately obtain the Hubble and deceleration parameters throughout our past light cone:

$$H(z) = H_0 \Big[\sum_j \Omega_{j0}(1 + z)^{3(1+w_j)} + \Omega_{k0}(1 + z)^2\Big]^{1/2}, \tag{5.47}$$

$$q(z) = \frac{1}{2} \sum_j \Omega_{j0}(1 + 3w_j)(1 + z)^{3(1+w_j)}. \tag{5.48}$$

With these parameters at hand, and armed with increasingly precise and abundant cosmological observations, we are then capable of adjusting and constraining our models, making precise (and independent) determinations of the parameters H_0, Ω_{k0} and Ω_{j0}. These, on their turn allow us to gravitationally infer the matter and energy content of the universe, and both contrast them with other more direct forms of observation (*e.g.*, the mater we *see* electromagnetically) and draw predictions/**retrodictions** about details of the evolution of the universe and their observational consequences. In what follows, we give a brief account of some of them, outlining the foundations and predictions of the standard (ΛCDM) cosmological model[G], which is extremely successful in *describing* the great majority of our cosmological observations up to this date[H].

Two particularly surprising results are that (i) the observed spatial curvature of our universe is essentially null (within experimental error) $\Omega_{k0} \sim 0$, and (ii) the greater part of energy density today is in the form of a nonobserved, extremely homogeneous energy form (which we call *Dark Energy*) with an equation of state $w \simeq -1$, resembling a cosmological constant Λ, its relative density being $\Omega_\Lambda \sim 0.7$. Furthermore, of the remaining 30% – which are virtually dominated by "dust" (nonrelativist, preussureless matter) and concentrate in known astrophysical structures (such as galaxies, clusters and superclusters) – only about 5% seem to correspond to baryonic matter, $\Omega_B \sim 0.05$. The remaining 25% correspond to some unknown species that (like dark energy) does not interact electromagnetically, thus called *(Cold) Dark Matter* (CDM), with $\Omega_{CDM} \sim 0.25$. For this reason, the standard cosmological model is also known as the ΛCDM model. Together, these two dominant (and so far uncomprehended) components form what we call *the dark sector*, comprising around 95% of all the energy in the observable universe today.

Now, analyzing how the universe was at earlier times, we find immediately from equation (5.46) that components with greater values of pressure (of w_j) become increasingly more significant as we look further in the past (at increasing values of z). Particularly, we see that Ω_Λ decreases in relative importance, coming to a shift where the universe was dominated by cold matter at about $z \sim 0.3$. Further, as we go back over 13 billion years, before matter clumped into galaxies and stars and planets could be formed, we come to a point where the universe was dominated by *radiation*, at about $z \sim 3600$; at this time, the temperature of the universe was extremely high (approximately $T \approx 3600 \times 2.7K \approx 10^4 K$ [I]). Temperatures and densities then become increasingly extreme into the radiation dominated era. If we extrapolate this era all the way back to an initial singularity, as predicted in a (radiation-dominated) FLRW model, we end up with what is called *the*

[G] For a more detailed account of all those points below, see (12–14).

[H] One very interesting exception that is arising in the most recent years is the so-called Hubble Constant Tension (44).

[I] The temperature of radiation scales proportionally to $a^{-1} \propto 1 + z$, and the temperature of the cosmic radiation that we observe today is about $2.7K$, as we shall discuss briefly.

Hot Big Bang scenario[J]. Given our knowledge from terrestrial experiments (specially at particle accelerators that reach very high energies), we can more or less safely extrapolate our predictions back to energies of about $kT \sim 10^4 GeV$, which correspond to extremely high redishifts $z \sim 10^{13}$ and, in the standard Hot Big Bang scenario (for which $a \propto t^{1/2}$ up until a singularity), to very early times: $t \sim 10^{-13} s$ after the Big Bang.

Our last direct observation window with electromagnetic radiation go back before any galaxies and stars were formed, at a redshift $z \sim 1100$ (roughly $t = 350.000$ years after the Big Bang), in the so-called surface of last scattering: at this time, the universe has undergone a phase transition, becoming so hot that nuclei and electrons cannot be bound together, and form an opaque plasma that constantly scatters photons. It is only after it cools enough for stable atoms to form that the universe became transparent, and this surface of last scattering forms an observable relic of the early universe, the so called Cosmic Microwave Background (CMB). We observe the CMB today as an extremely isotropic radiation from every direction in the background sky, with a distribution that fits extremely well that of a blackbody with a temperature of about $2.73K$ (as it has redshifted for a factor of more than one thousand since it was emmited in a hot plasma). If we correct for a dipole anysotropy (which is attributed to the peciliar velocity of the earth with respect to a cosmic isotropic worldline), we end up with a very homogeneous temperature distribution in it, with very small relative fluctuations $\frac{\delta T}{T} \sim 10^{-5}$. The universe was indeed extremely homogeneous at those early times.

In the next section, then, we begin to investigate the question of why it was so homogeneous (and spatially flat) to start with. Along with the Dark Sector, these are two of the greatest open questions of modern cosmology.

5.1.3 Fundamental problems in the ΛCDM model

The universe that we observe today, of course, is far richer and more complex than a perfectly homogeneous and isotropic spacetime. Although it is roughly homogeneous and isotropic on very large scales, it becomes richly filled with strutures of various sizes and types as we dwell in smaller ones. For the roughly thirteen billon years that have transcurred since decoupling, matter has been collapsing gravitationally, forming the various structures that we see today, ranging through superclusters, clusters, galaxies, and down to stellar systems and individual celestial bodies like stars and planets. Although the existence of these structures is a commonplace from our perspective as inhabitants of this universe, their formation process turns out to require a very particular adjusting of

[J] In fact, one does not really require the very singular hyphothesis of a perfectly spatially homogeneous and isotropic universe. It was originally shown by Penrose (and then applied by Hawking in a cosmological context) that, according to General Relativity, singularities should actually form under much more generic conditions; see, for instance, (5, 25) for singularity theorems.

cosmological parameters, as we can infer in the light of our cosmological models.

On the one hand, they require initial fluctuations in density, which act as seeds for gravitational collapse; these turn out to be given precisely by the tiny fluctuations whose imprint we observe in the CMB today. On the other, they require that the average energy density of the universe is extremely well tuned to the critical density $\Omega \sim 1$, $|\Omega_k| \ll 1$, so that curvature does not quickly dominate over matter, and either recollapses the universe (if $\Omega_k < 0$) or makes it expand too fast for structures to form (if $\Omega_k > 0$). This is known as the flatness problem in cosmology.

To analyze this problem more, let us consider the spatially spherical case ($k = 1, \Omega_k < 0$) and estimate the recollapsing time of the universe for a small imbalance in Ω_k at some given initial time. This leaves us with the question of what would be a reasonable time to impose "initial conditions" in the universe. It certainly cannot be at $t = 0$, as there should be a singularity there (according to the standard Hot Big Bang model), and, classically, any finite time seems equally arbitrary to impose them. We shall argue in more detail in section 5.3.1 that a relatively natural time to do so should be given by the Planck time $t_p \sim 10^{-43}s$. For the time being, we assume this to be true, and evaluate the evolution of $\Omega_k(t)$ starting from a small imbalance $\Omega_{kp} < 0$ (here, all subscripts p refer to quantities evaluated at $t = t_p$) :

$$\Omega_k(t) = 1 - \Omega(t) = \frac{H_p^2 a_p^2}{H^2(t)a^2(t)}(1 - \Omega_p). \tag{5.49}$$

One can see in the Friedmann equations (5.11) that the turnpoint between expansion and contraction happens when the matter and curvature terms cancel each other out $\Omega_k = -\Omega$. Since H will vanish at that point, it actually must happen for $\Omega \to \infty$ and $\Omega_k \to -\infty$. Before that, during the time that matter is considerably dominant $\Omega \sim 1, |\Omega_k| \ll 1$, considering a radiation-dominated early universe, we have:

$$H^2(t) \simeq \frac{8\pi}{3}\rho_p\left(\frac{a_p}{a(t)}\right)^4. \tag{5.50}$$

Since for an approximately flat, radiation-dominated universe, $a \propto t^{1/2}$, that yields:

$$\Omega_k(t) = 1 - \Omega(t) \simeq (1 - \Omega_p)\frac{t}{t_p} = \Omega_{kp}\frac{t}{t_p}. \tag{5.51}$$

If we then estimate "half the age of the universe" ($T_U/2$) by a time when Ω_k and Ω become comparable, say, at $\Omega_k \sim -1$, and extrapolate (5.51) up to that point, we obtain:

$$\Omega_{kp} \sim -\frac{t_p}{T/2}. \tag{5.52}$$

Thus, in order that the universe does not recolapse within a very short time T, one must tune the energy density at t_p extremely close to critical density. For concreteness, we calculate the upper limits for this imbalance for a few values of T:

$$\begin{cases} T = 10^{-36}s : |\Omega_{kp}| \simeq \frac{\Omega_p - 1}{\Omega_p} \lesssim 5 \times 10^{-8}, \\ T = 10^{-26}s : |\Omega_{kp}| \simeq \frac{\Omega_p - 1}{\Omega_p} \lesssim 5 \times 10^{-18}, \\ T = 10^{18}s : |\Omega_{kp}| \simeq \frac{\Omega_p - 1}{\Omega_p} \lesssim 5 \times 10^{-42}. \end{cases} \tag{5.53}$$

Particularly, the last case corresponds to the present age of our universe (which is not nearly in a process of recollapse), from which we see that a truly extreme fine-tuning in the curvature would be necessary for the universe to still exist up to this day. On the other hand, if we had a positive initial imbalance in Ω_{kp}, the universe would have expanded drastically faster than it did, not allowing for the formation of any of the structures that we oberve today.

Having noted what seems to be an extreme coincidence regarding the spatial curvature of our universe, we turn our attention to what appears to be another great coincidence: why was the early universe so spatially homogeneous to start with? If we want to avoid the extreme coincidence of merely postulating that it was "born" homogeneous (but still with tiny fluctuations that allowed the formation of structures), a reasonable hypothesis would be that it had time to evolve into an equilibrium temperature and density configuration, so that our first observations actually measure this equilibrium profile. However, a problem that emerges in the standard Hot Big Bang scenario is that there are generally *particle horizons.* Those are causal horizons in the past, due to the fact that each isotropic worldline has not (for a radiation-dominated expansion) had time to be in causal contact with all other isotropic worldlines, and thus come to equilibrium with them. This is known as the Horizon Problem.

To take a closer look at this problem, it is useful to consider a spatially flat FLRW spacetime, and cast its metric in a confomally Minkowskian form:

$$ds^2 = a^2(\eta)\Big[d\eta^2 - d\mathbf{x}^2\Big], \qquad \eta_0 - \eta = \int_t^{t_0} \frac{dt'}{a(t')}. \tag{5.54}$$

Recall that, as conformal transformations preserve the light cones, conformally related spacetimes will have the same causal structure. Thus, this FLRW spacetime will have the same causal structure as (a portion of) Minkowski spacetime. Note then that, in order for all isotropic worldlines to be causally connected at a time t_0 (a conformal time η_0), it is necessary that η extends all the way past to $-\infty$ when $t \to 0$; otherwise, this spacetime will be conformally related to just a portion of Minkowski spacetime for which $\eta > \eta_{sing}$, given by:

$$\eta_{sing} \equiv \eta_0 - \int_0^{t_0} \frac{dt'}{a(t')}. \tag{5.55}$$

(The value for which we define η_0 here is arbitrary and irrelevant. The point is whether η_{sing} will be a finite time or extend all the way back to $-\infty$.)

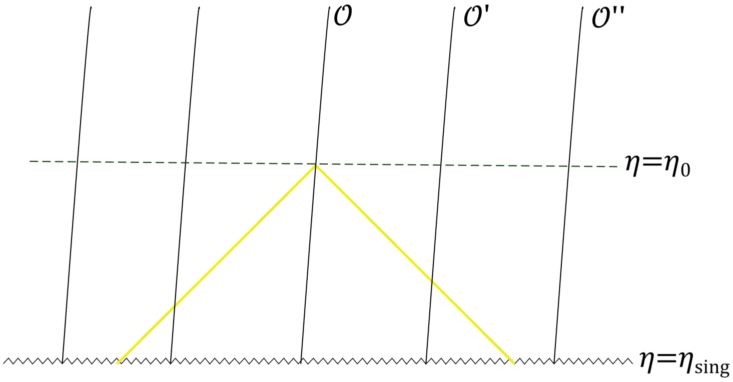

Figure 44 – Conformally flat FLRW spacetime, represented in conformal coordinates. If η_{sing} turns out finite, this space will only be conformally related to a portion of Minkowski spacetime, and distant observers will not have had time to come in causal contact. Tracing the past light cone of an observer $\mathcal{O}$ at the time η_0, we see that it intersects some of the other worldline (like $\mathcal{O}'$), but not sufficiently distant ones (like $\mathcal{O}''$).
Source: By the author.

It is apparent from (5.55) that for a power-law exapansion $a \propto t^\lambda$, with $0 < \lambda < 1$, there will be causal horizons. In fact, such horizons will generally occur for any monotonically *decelerated* expansions. In such cases, the Hubble radius $H^{-1}(t)$ of the universe will *increase* with time, and more isotropic worldlines will become causally connected as time passes (that is, increasingly more observers will have had time to interact with one another)[K].

To analyze the distances across which different points in the far sky were connected at past time t, it is useful to evaluate an instanteneous causal radius around an arbitrary isotropic worldline as a function of t (particularly, we will be interested at the time t_d when

[K] In contrast, for *accelerated* expansions, there will be future causal horizons (*i.e.* event horizons) and 2 isotropic observers who were causally connected at early times will later cease to be. This is quite clearly illustrated in the example of an exponentially expanding (de Sitter) space, which will be discussed in section 5.3.

the CMB was formed). We estimate this causal radius for a dust-dominated universe[L] ($a \propto t^{2/3}$):

$$d_G(t) = ca(t)\int_0^t \frac{dt'}{a(t')} = 3ct \tag{5.56}$$

(where we have reincorporated $c \neq 1$ into the formulas to convert to usual distance values).

Thus, for our present universe $t = t_0 \sim H_0^{-1} \sim 10^{10}Yr$, we have:

$$d_G(t_0) \sim 9 \times 10^3 Mpc. \tag{5.57}$$

Then, taking the time of matter-radiation decoupling t_d as approximately given by:

$$\frac{a_0}{a_d} \simeq \left(\frac{t_0}{t_d}\right)^{2/3} \sim 10^3 \qquad \Rightarrow \qquad t_d \sim 3 \times 10^5 Yr, \tag{5.58}$$

we may estimate the causal radius at the formation of the CMB:

$$d_G(t_d) \simeq 3ct_d \sim 0.3Mpc\,. \tag{5.59}$$

This represents a geometrical distance on Σ_{t_d}. We can then calculate the corresponding geometrical distance D at the present time[M] t_0 (*i.e.*, in Σ_{t_0}) by streching it by a factor of $\frac{a_0}{a_d} = \left(\frac{t_0}{t_d}\right)^{2/3}$:

$$D = \left(\frac{t_0}{t_d}\right)^{2/3} d_{light}(t_d) \sim 300Mpc\,. \tag{5.60}$$

Thus, the archlenght in the background sky corresponding to 2 causally connected points should be roughly around $D \sim 300Mpc$. This arc, on its turn, should lie in a sphere of radius $R = d_{light}(t_0) \sim 9 \times 10^3 Mpc$, such that it will correspond to an angle θ roughly given by:

$$\theta = \frac{D}{R} \sim \frac{300Mpc}{9000Mpc} = \frac{1}{30}\text{radians} \sim 2^\circ. \tag{5.61}$$

[L] Note that, indeed, the universe was dust-dominated for most of the time up until the matter-radiation decoupling in the CMB. Furthermore, it makes little difference if we consider $\lambda = 2/3$ or $\lambda = 1/2$ in this calculation.

[M] This stretctched distance will be observationally meaningful because we are only interested in calculating an angular width from it, and angles are preserved in an isotropic expansion.

This means that, according to the standard Hot Big Bang scenario, patches of the CMB separated by more than approximately 2° should be causally disconnected! Thus, the fact that the we observe an extremely isotropic CMB would indeed be entirely coincidental, as it could not be attributed to any causal process of thermalization.

For completeness, we also mention the issue of topological defects in grand unification theories (GUTS). Based on the successful unification of the electromagnetic and weak interactions in the so-called electroweak force, at energy scales of about $\sim 1TeV$, it has been proposed that the electroweak and strong forces should become unified at energy scales many orders of magnitude grater, of about $10^{12}TeV$, corresponding to temperatures of around $10^{28}K$. As we have mentioned in the last section, extrapolation of our cosmological model predicts that the universe should indeed have risen to arbitrarily high temperatures, reaching the ones corresponding to grand unification at about $t = 10^{-36}s$. The issue when we try to combine these theories with our cosmological model is that, at the point the universe cools enough to undergo a phase transition from a grand unified force to separated strong and electroweak ones, it is expected that topological defects – particularly, very massive particles (monopoles), with $m_M \sim kT_{GUT} \sim 10^{12}TeV$ – are produced in a certain abundance; roughly, we can estimate there to be one per causal sphere of radius $d_G(t_{GUT}) = 2t_{GUT}$ (for a radiation-dominated universe, $a \propto t^{1/2}$), such that their numerical density and mass density should have been:

$$n_M(t_{GUT}) \sim (2t_{GUT})^{-3}, \qquad \rho_M \sim \frac{m_M}{(2t_{GUT})^3} \sim 10^{94}TeV/m^3. \tag{5.62}$$

Although astoundingly high, this density should have been relatively insignificant compared to the radiation-dominated critical density at the time. A quick estimate in the standard Hot Big Bang scenario yields:

$$\rho_{rad}(t_{GUT}) \sim 10^{104}TeV/m^3 \qquad \Leftrightarrow \left.\frac{\rho_M}{\rho_{rad}}\right|_{t_{GUT}} \sim 10^{-10}. \tag{5.63}$$

However, since these monopoles would have been extremely heavy, they should behave as noninteracting "dust" since very early times, and would become quickly dominating over radiation (the latter was dominant in the very early universe up to about $10^{12}s$). Making use of (5.45), we would estimate for our present universe:

$$\frac{\Omega_M(t_0)}{\Omega_{rad}(t_0)} \simeq 10^{18}, \tag{5.64}$$

and, since radiation today is about only 10^{-5} of the critical density of the universe, we would have:

$$\Omega_{M0} \equiv \frac{\rho_{M0}}{\rho_{c0}} \sim 10^{13}, \tag{5.65}$$

in a screaming contradiction with our observations of $\Omega_{total} \sim 1$.

Summarizing, although the ΛCDM model is extremely successful in decribing our observations with very few parameters, it turns out to carry quite significant fundamental issues, either on its own or in conjunction with other physical theories (albeit, in non-tested regimes in the case of monopoles in GUTs). How, then, could we handle all of these issues? Well, it so happens that a single (although considerably long shot) solution to all of them can be found by postulating a very short primordial inflationary period lasting until about $\sim 10^{-33}s$ during which the universe would have expanded quasiexponentially by a factor of at least $\sim 10^{26}$ times. In the next sections, we shall explore the foundations and developments of this quite extreme proposal, both illuminating how it could be realized through field theory in curved spaces, and attempting to draw its connections to observational quantities.

5.2 Field Theory and Inflation; spontaneous symmetry breaking

Before we actually dwell in the cosmological developments of inflation, we shall take a moment to describe how a finite inflationary phase could emerge in the joint dynamics of spacetime and matter fields in the first place. We have already seen in the previous chapter that the renormalized vacuum energy of noninteracting quantized fields in curved (de Sitter) spacetimes could give rise to a contribution in the form of cosmological constant – whose gravitational effects, when taken into account, should ultimately produce an (eternally) exponentially expanding universe.

Then, to be able to grasp how vacuum energy could give rise to a finite-lasting inflationary phase, we are forced to extend our analysis beyond free fields and consider interacting ones. Given the difficulties in fully analyzing quantized interacting fields in curved spacetimes[N] and the fact that many relevant features and effects of inflationary cosmology already arise in a classical regime, we will now turn our attention back to classical fields in curved spacetime. Later, we shall consider a perturbative approach to quantize linearized fluctuations of our field.

Throughout the rest of this chapter we shall consider, for concreteness and simplicity, a single self-interacting scalar field ϕ, minimally coupled to gravity, whose Lagrangian will be generically of the form:

[N] For a further discussion on interacting fields, particular on the self-interacting $\lambda\phi^4$ model, see the final chapter of (1) and section 6.7 of (2).

$$\mathscr{L} = \frac{1}{2}(\nabla^{\mu}\phi)(\nabla_{\mu}\phi) - V(\phi), \tag{5.66}$$

where $V(\phi)$ is a generic potential.

We can immediately derive the Euler-Lagrange equation for this field, which reads:

$$\Box\phi + V'(\phi) = 0, \tag{5.67}$$

where $V' = \frac{dV}{d\phi}$.

Clearly, for the usual harmonic potential $V(\phi) = \frac{m^2}{2}\phi^2$, $m^2 \geq 0$ we recover a free scalar field, which can be decomposed in an infinite collection of decoupled harmonic oscillators, with positive and negative frequency modes $u_{\mathbf{k}} \propto e^{\mp ik_{\mu}x^{\mu}} = e^{\mp i\omega t}e^{\pm i\mathbf{k}\,\cdot\,\mathbf{x}}$, being ω and $\mathbf{k}$ real. However, if we make a signal inversion, considering a mass $\mu^2 = -m^2 < 0$, we end up with the well-known issues for a theory with a Hamiltonian unbounded from below: there will be no ground state and the dynamics is rendered unstable by the presence of arbitrarily negative energy eigenvalues. Indeed, in a such a case we have the field equations:

$$\left[\Box - m^2\right]\phi = 0, \tag{5.68}$$

for which we still have modes of the kind $u_{\mathbf{k}} \propto e^{\mp ik_{\mu}x^{\mu}}$, but which will have a dispersion relation:

$$k_{\mu}k^{\mu} = (k^0)^2 - \mathbf{k}^2 = -m^2, \tag{5.69}$$

such that, for $|\mathbf{k}| < m$, we are forced to consider imaginary frequencies[O] $\omega = \pm k^0 = \pm i\sqrt{m^2 - \mathbf{k}^2}$, which lead to (unbounded) exponentially growing modes:

$$u_{\mathbf{k}} \propto e^{\pm\sqrt{m^2-\mathbf{k}^2}\,t}e^{\pm i\mathbf{k}\,\cdot\,\mathbf{x}}. \tag{5.70}$$

Of course, such a plainly pathological model is of little use to us on its own. However, it is useful to transparently reveal the instabilities of a vacuum state surrounding a local

[O] In a simmilar spirit, one could also consider imaginary wave vectors $\mathbf{k}$ (which also result in imaginary frequencies). These should be generally forbidden in well-behaved theories so that we do not end up with spatially exponentially divergent modes, but one is forced to consider them if space and time are to be treated in an equal-footing. The asymmetry here arises in the way we split our modes to satisfy an initial condition and boundary problem.

maximum, which we will explore in better-behaved model. A very interesting nontrivial potential with global minima is the so called $\lambda\phi^4$ model, whose potential is given by:

$$V(\phi) = V_0 - \frac{m^2}{2}\phi^2 + \frac{\lambda}{4}\phi^4. \tag{5.71}$$

This potential has two symmetrical global minima (corresponding to two stable vacua) at $\phi = \phi_\pm \equiv \pm m/\sqrt{\lambda}$ and a local maximum (corresponding to an unstable vacuum) at $\phi = 0$. This relatively simple model turns out to yield a very rich structure: as in the case of a repulsive oscillator (with $\mu^2 < 0$), it will have exponential instabilities for modes around the local maximum at $\phi = 0$; further, if the energy of this field falls below V_0, it will be classically confined at one of the potential wells, either around ϕ_+ or ϕ_-, which results in a spontaneous symmetry breaking. Particularly, if the field has an energy very close to its absolute minimum, it will have nearly constant values, just making small oscillations around one of its stable vacuum states: $\phi = \phi_\pm + \delta\phi$, $|\delta\phi| \ll |\phi_\pm| = m/\sqrt{\lambda}$. This last situation is particularly interesting, because it allows one to break the field in a constant classical contribution plus perturbations, which one can quantize in a linearized regime. Considering, for example, the vacuum at $\phi_0 = \phi_+$, we obtain:

$$\begin{aligned}\mathscr{L} =& \frac{1}{2}(\partial_\mu\delta\phi)(\partial^\mu\delta\phi) - \frac{1}{2}(3\lambda\phi_0^2 - m^2)(\delta\phi)^2 - \lambda\phi_0(\delta\phi)^3 - \frac{\lambda}{4}(\delta\phi)^4 \\ &+ \phi_0(m^2 - \lambda\phi_0^2)\delta\phi + \left(\frac{m^2}{2}\phi_0^2 - \frac{\lambda}{4}\phi_0^4\right).\end{aligned} \tag{5.72}$$

Here, the linear term in $\delta\phi$ will vanish because $\phi_0^2 = \frac{m^2}{\lambda}$, and the constant term can be ignored for quantization purposes (the only nontrivial role that this term may play is gravitational). Then, neglecting terms of cubic order or higher for $\delta\phi$, we can quantize as a field of effective quadratic mass $\mu^2 = 3\lambda\phi_0^2 - m^2 = 2m^2$.

A further aspect, which would have significant cosmological consequences is the possibility that topological defects may arise. At high energies the field can symmetrically explore configurations around both ϕ_+ and ϕ_-. As it goes into lower energies however, it is forced to collapse in either one of these regions, spontaneously breaking its symmetry. Over very large, causally disconnected regions, however, one has no reason to expect that the field will uniformly collapse in the same region (either ϕ_+ or ϕ_-); more realistically, it should form domains in which ϕ has decayed in either value. Then between any two domains, there must be a transitioning region where the field has intermediate values, which will have very high energy densities: such regions are called domain walls. Since this is a simple scalar field, these topological defects between different vacua will be quasi 2-dimensional. In more realistic theories, with different groups of symmetries, the defects may be 1-dimensional (cosmic strings) or 0-dimensional (monopoles).

Moreover, if ϕ happens to be interacting with other fields in nature, this spontaneously broken classical value may give rise to a mass term for the latter fields[P]. For example, if one considers a massless Dirac field ψ coupled to our scalar field by an interaction term $\mathscr{L}_I = -h\phi\bar{\psi}\psi$ (h being a coupling constant), we have a total Lagrangian:

$$\begin{aligned}\mathscr{L} &= \mathscr{L}_\phi + \mathscr{L}_\psi + \mathscr{L}_I \\ &= \frac{1}{2}(\partial_\mu\phi)(\partial^\mu\phi) + \frac{m^2}{2}\phi^2 - \frac{\lambda}{4}\phi^4 + \bar{\psi}i\gamma^\mu\partial_\mu\psi - h\phi\bar{\psi}\psi \end{aligned} \tag{5.73}$$

(where, for simplicity, we are considering spacetime to be flat).

Then, again, breaking ϕ as $\phi_0 + \delta\phi$ around a minimum at ϕ_0, we obtain:

$$\mathscr{L} = \frac{1}{2}(\partial_\mu\delta\phi)(\partial^\mu\delta\phi) + \frac{\mu^2}{2}(\delta\phi)^2 + \bar{\psi}(i\gamma^\mu\partial_\mu - h\phi_0)\psi - h\bar{\psi}\psi\delta\phi + \mathcal{O}\Big((\delta\phi)^3\Big) + cte, \tag{5.74}$$

and, comparing the third term with the massive Dirac Lagrangian, we clearly find that a mass term for the fermions emerges: $m_\psi = h\phi_0 = hm/\sqrt{\lambda}$.

Similarly, in scalar electrodynamics, where one couples the electromagnetic field to a complex scalar field[Q] χ, one can build the so-called Abelian Higgs model:

$$\mathscr{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + D_\mu\chi^* D^\mu\chi + \mu^2\chi^*\chi - \lambda(\chi^*\chi)^2, \tag{5.75}$$

being $F_{\mu\nu} = 2\partial_{[\mu}A_{\nu]}$ the Faraday tensor, and D_μ the Gauge covariant derivative, which acts on charged scalar fields as $D_\mu\chi = (\partial_\mu - ieA_\mu)\chi$ (and $D_\mu\chi^* = (\partial_\mu + ieA_\mu)\chi^*$). This obeys a local gauge symmetry:

$$\begin{cases} A_\mu(x) \to A'_\mu(x) = A_\mu(x) + \partial_\mu\xi(x), & (5.76\text{a}) \\ \chi(x) \to \chi'(x) = \chi(x)e^{ie\xi(x)}, & (5.76\text{b}) \end{cases}$$

which one can exploit to make χ real everywhere. If $\chi(x) = \rho(x)e^{i\theta(x)}$, we can adjust the gauge making $\xi(x) = -\theta(x)$, so that $\chi(x) \to \chi'(x) = \chi(x)e^{-i\theta(x)} \equiv \frac{1}{\sqrt{2}}\varphi(x)$. In this transformed gauge, the Lagrangian reads:

[P] It is beyond the scope of the present work to thoroughly present the exciting topics of symmetry breaking either in wide generality or in its applications to the fields of High Energy Physics (HEP) and Cosmology. Still, the author feels strongly compelled to make a brief discussion in some qualitative aspects of this subject, as they bear an intimate relation with many topics in this text. For a more thorough discussion, the reader is referred to the later chapters of (6) for a more pedestrian introduction to this topic in HEP, to (4) for a further exposition in a cosmological context (including molopoles and more general topological defects in the universe), or to (17) for a more mathematically rigorous, model-indepent presentation.

[Q] It is necessary that the field is complex so that Noether Charge will emerge which is associated to gauge transformations, and thus can be identified with electrical charge.

$$\mathscr{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + (\partial_\mu\varphi)(\partial^\mu\varphi) + \frac{\mu^2}{2}\varphi^2 - \frac{\lambda}{4}\varphi^4 + \frac{e^2}{2}\varphi^2 A'_\mu A'^\mu. \tag{5.77}$$

Then, if we once again split $\varphi = \phi_0 + \delta\varphi$ in a minimum of its potential ($|\varphi_0| = \mu/\sqrt{\lambda}$), we clearly end up with a mass term for the electromagnetic field A^μ:

$$m_A^2 = -e^2\varphi_0^2 = -\frac{e^2\mu^2}{\lambda}. \tag{5.78}$$

These simple models exemplify (at a merely qualitative level, in our superficial exposition) the types of phenomena that interacting field theories can describe, and are suggestive of the types of phase transition the universe may have undergone during a inflationary period. However, for a description of inflation itself, that is, of a finite period during which the universe has expanded at extremely fast and accelerated rate, we shall focus our attention solely on the scalar field ϕ, and describe its evolution within a single domain.

In our $\lambda\phi^4$ model, it is not hard to conceive how one could obtain such a finite inflationary phase: if we adjust the potential constant V_0 in (5.66) such that $V(\phi_\pm) \approx 0$ (and $V(\phi = 0) \approx V_0 > 0,\ V'(\phi = 0) = 0$), it could spend a considerable amount of time near its unstable local maximum, driving a nearly constant relative expansion (*i.e.*, a quasiexponential expansion) and subsequently decay into one of its local minima, giving up energy to ordinary forms of matter and energy; after it decayed, inflation would cease and the expansion would become dominated by other energy forms. In the next section, we shall then concretely consider an inflationary scenario, and derive a more quantitative description of this quasiexponential expansion dynamics driven by a classical scalar field.

5.3 Inflationary Cosmology: the chaotic inflation scenario

Historically, there have been a number of proposals for inflationary models and scenarios[R], all of which attempted to address the same basic issues raised in the last section by proposing some mechanism through which the universe might have undergone an extremely abrupt period of accelerated expansion in its very early history. Despite these basic similarities, specific models vary widely both in their qualitative features and in their quantitative predictions, which may range as much as dozens of orders of magnitude in several quantities (such as duration of inflation and reheating, magnitude of vacuum energy, magnitude and spectrum of fluctuations, types and abundance of topological

[R] A very complete and yet concise account of the history and successive developments of these models and scenarios can be found in section 1.6 of (4) (and see the various references therein for particular models and developments).

defects, etc.), many of which are so far very loosely constrainable by observations[S]. The specific scenario that we shall present here, namely, the chaotic inflation scenario, has many of its developments due to Andrei Linde, and it is more thoroughly presented along with other inflationary models on his book (4), which is the main reference for our present exposition. This scenario not only allows us to solve the motivating issues that pushed us towards inflation in the first place, but it also provides a somewhat 'natural' background to the discussion of initial conditions and to the grand questions of why our observable universe has the form it has, as well as a reasonable framework to analyze the primordial fluctiations in the CMB.

For clarity of ideas and computability, we shall restrict the present exposition of this scenario to the case of a real scalar field, with Lagrangian (5.66); concretely, the reader may often bear in mind the $\lambda\phi^4$ model, although we shall often make estimates considering different power-law potentials, taking the form:

$$V(\phi) = \frac{\lambda_n \phi^n}{n M_p^{n-4}}, \tag{5.79}$$

being $n > 0$, M_p the Planck mass and $0 < \lambda_n \ll M_p^{n-4}$. This encompasses the harmonic and quartic potentials (where we make the identifications $m^2 = \lambda_2 M_p^2$ and $\lambda = \lambda_4$).

5.3.1 Initial conditions

Let us now turn our attention to the question of initial conditions. In our previous analysis we have discurred about the great coincidence it would have been to have such a homogeneous universe in the time of decoupling had the universe expanded dominated by radiation ($a \propto t^{1/2}$) all the way back to an initial singularity. Furthermore, it remained an open question why its fluctuations were so small, having a typical relative amplitude of 10^{-5} (rather than of any other conceivable value). The problem became even more acute when we noticed how particular the values of some parameters must have been (particularly, how spatial curvatured must have been extremely fine-tuned near 0) so that our universe could evolve and form structures in the way it did (and, among other things allow for the emergence of life).

In a first sight, when we consider a primordial inflationary scenario, it seems that we can do little better than to push the same problem back in time, reaching an instant for which we must specify *some* initial condition (which should in principle not escape the same issues of fine-tuning to allow the realization of *our* observed universe). We shall argue, however, that one may actually obtain an observable universe such as ours

[S] Nevertheless, it is noteworthy how the observational precision has impressively improved in the last two decades, allowing for progressively better constraints. For a contemporary account of the state-of-the-art and perspectives for such measurements, see *e.g.* (41).

from fairly generic initial conditions. More precise, one may obtain *many realizations* of patches of the universe that look like a FLRW spacetime over extremely large scales – many orders of magnitude larger than our observable universe –, such that it would be reasonable to assume that *somewhere* there would be a patch such as ours. In a loose analogy with biological evolution, the universe would not have had to have developed in very particular way to end up with very special and complex structures – the mere fact that it could randomly explore a huge (virtually infinite) sample of different configurations in a sufficiently large space would assure that *some* of these configurations could result in a very special patch of spacetime. Of course, since such a hypothesis refers to regions much larger than our observable universe, many of its fundamental consequences can be no more than unobservable conjectures[T]. However, this does not mean that an inflationary scenario based on it will be devoid of observational consequences *in our observable universe*. Our analysis, then, shall be primarily concerned with the latter (although we would like to stress that a line between one and the other is by no means sharp or *a priori* obvious, and that much can be gained in the latter by pursuing and exploring one's ideas with a good degree of open-mindness, beyond what is obviously verifiable).

Then, to figure more precisely what would be reasonable to postulate as initial conditions for our model, we recur to what we *know* to be the conditions of our observable universe in the very early past. We know that at a point earlier than the formation of the CMB the universe was very hot and came to be dominated by relativistic degrees of freedom (particularly, electromagnetic radiation), contracting back to the past at a rate $a \propto t^{1/2}$ for many orders of magnitude of a. We can extrapolate this radiation dominated expansion at least all the way back through primordial nucleosynthesis (first minutes), electroweak unification ($\sim 10^{-10}s$) and even somewhat before that (physics at known energy scales allows us to safely extrapolate back to about $\sim 10^{-13}s$). The standard Hot Big Bang Scenario amounts to assuming that such a rate of expansion extrapolates *all the way back to* $t = 0$, where there would be a primordial singularity. However, even if we did not have any of the aformentioned issues that motivate us to postulate an early inflationary phase, one fundamental difficulty remains in extrapolating this analysis back into arbitrarily high curvature and energy scales: quantum gravity. When we reach curvatures as high as the Planck scale, a dynamical description of the universe (including geometry and matter) in terms of a classical continuous spacetime (and matter fields propagating in it) is no longer expected to hold.

Since our description of inflation must apply only in times much earlier than those to which we can safely extrapolate with experimentally verified theories, but yet late enough so that it should be prone to a description in terms or classical spacetimes, we try to impose them at 'the edge' of a classical descripton, at Planck time t_p. A rather

[T] Albeit, in the author's opinion, they make for a quite appealing picture compared to the fine-tuning alternative.

simple, and somewhat physically motivated possibility (given the known homogeneity of the early universe) would be start the inflaton field ϕ in a constant (homogeneous) value ϕ_0 corresponding to a vacuum state. However, we can immediately see that such a condition would be even more singular than a FLRW space (almost) perfectly homogeneous, and that it further presents difficulties of consistency at a quantum level (it cannot take quantum fluctuations into account), and of compatibility with the small inhomogeneities in the early universe, which appear in the CMB and are crucial for the formation of structure in our universe.

In fact, regardless of what the initial state and primordial dynamics of the universe were (and particularly, whether that dynamics depended solely or primarily on a scalar field ϕ), the energy density of the universe cannot be determined with a precision greater than $M_p l_p^{-3}$, in virtue of the uncertainty principle (here, we shall associate inverse length scales with mass scales, and generally make reference just to the Planck mass – then we say this energy density cannot be determined with a precision greater than M_p^4). Thus, instead of assuming this extremely specific (and ultimately inadequade) initial condition, we make more generic (and physically reasonable) hypotheses that, for $t \sim M_p^{-1}$:

$$(\partial_0\phi)(\partial^0\phi),\, (\partial_i\phi)(\partial^i\phi),\, V(\phi) \lesssim M_p^4, \tag{5.80}$$

and similarly for any scalars derived from spacetime curvature:

$$R^2, R_{ab}R^{ab}, R_{abcd}R^{abcd} \lesssim M_p^4, \quad R_a{}^b R_b{}^c R_c{}^a, R_{abcd}R^{ac}R^{bd} \lesssim M_p^6, \quad \text{etc.} \tag{5.81}$$

We stress that, in fact, it is precisely 'after the instant' in which condition (5.81) is satisfied that we may even speak about a dynamical description in a classical spacetime, and specify an initial condition for ϕ.

One may then argue that, within these (consistency-binding) intervals given by equations (5.80) and (5.81), there is no *a priori* reason to expect that $(\partial_\mu\phi)(\partial^\mu\phi) \ll M_p^4$, $V(\phi) \ll M_p^4$, or $R^2 \ll M_p^4$ (more precisely, given our ignorance about these initial conditions – and the physics that governs them – it should be more or less equally plausible to expect *any* values within physically reasonable restrictions), so that it would be more likely to assume that these quantities had initial conditions with values of the order (*i.e., not much smaller than*):

$$(\partial_0\phi)(\partial^0\phi) \sim (\partial_i\phi)(\partial^i\phi) \sim M_p^4, \tag{5.82a}$$

$$V(\phi) \sim M_p^4, \tag{5.82b}$$

$$R^2 \sim M_p^4 \tag{5.82c}$$

(where (5.82c) should be understood in the same sense as eq. (5.81), with "R^2" standing for any quadratic scalars built from curvature, and all analogous equations with appropriate dimensions for different powers of R).

For the time being, we shall assume that the initial conditions are indeed given approximately by eqs. (5.82) and, in the subsequent discussion, we will try to explore and better understand the implications of this assumption.

5.3.2 Quasiexponential expansion and *slow-roll* inflation

The treatment of the evolution of the universe under these 'generic' initial conditions is still an extremely complicated task. However, motivated by the known conditions of our observable universe and by our requirements for inflation, we are hinted to turning our attention to particularly symplyfing circumstances, namely, *to portions of the universe in which the dynamics yield an approximatly exponentially expanding FLRW universe (i.e. a de Sitter space).*

A very important symplifying feature of de Sitter spaces is that they have an event horizon at a radius H^{-1} surrounding each isotropic observer $\mathcal{O}$, and that all other isotropic worldlines eventually fall outside this horizon, losing causal contact with $\mathcal{O}$. Similarly to black holes, de Sitter spaces obey so-called "no-hair" theorems, which ultimately imply that any effects due to matter and energy that fall outside the horizon of a given domain will be exponentially dampened out and no longer affect the dynamics inside this domain; thus, any spacetimes that are locally approximately like a de Sitter space (whose total stress tensor obeys $T_{ab} \approx \Lambda g_{ab}$) for a large enough region will exponentially approach a de Sitter space. For such a behaviour to be realizable, the domain in which an approximately exponential expansion happens must be bigger than $\sim 2H^{-1}$ (*i.e.* the diameter of a Hubble sphere); as we shall see briefly, this will correspond to the domain for which T_{ab} is dominated by the potential term $V(\phi) \approx cte$. Well, in the 'instant' that $V(\phi) \sim M_p$, the Hubble radius will actually be as small as it can possibly be (to be described classically): of the order $H^{-1} \sim M_p^{-1}$.

Conversely, it should be indeed necessary that the expansion in such patches is approximately exponential for the horizon (located at a *dynamical* radius $H^{-1}(t)$) to recede at a sufficiently slow rate so that the primordial inhomogeneities can fall out of it and cease to affect the dynamics inside the relevant causal domain. A rough estimate for this required slowness may be obtained by noting that the recession velocity *of an object located at the horizon* should be of the order $\sim H^{-1}H = 1$ (by a simple application of the Hubble law (5.20) for $R \approx H^{-1}$) whereas the recession velocity *of the horizon itself* is $|\frac{d}{dt}H^{-1}| = \dot{H}H^{-2}$. Then this condition will be satisfied if $\dot{H} \ll H^2$.

We may then conclude that, in order for inflationary regions to emerge near the Planck epoch with initial conditions (5.82) (and subsequently grow to considerable sizes), it should be enough that they occur in *any* region of the universe with a minimal size liable to a description in terms of classical spacetime: $l \sim M_p^{-1}$.

We point out here that a particular consequence of condition (5.82b) (for potentials obeying the condition (5.79)) will be that initial values ϕ_0 of the field are tipically very large ($\gg M_p^4$), so that possible variations in a causally relevant scale should be comparatively small. For example, for the power-law potentials (5.79) with $n = 2, 4$, $V(\phi_0) \sim M_p^4$, we will have that:

$$\begin{cases} V(\phi) = \dfrac{m^2}{2}\,\phi^2, \text{ with } m \ll M_p \\[2ex] V(\phi) = \dfrac{\lambda}{4}\,\phi^4, \text{ with } \lambda \ll 1 \end{cases} \quad \Longrightarrow \quad \begin{cases} \phi_0 \sim \dfrac{M_p}{m} M_p \gg M_p, & \text{(5.83a)} \\[2ex] \phi_0 \sim \lambda^{-\frac{1}{4}} M_p \gg M_p, & \text{(5.83b)} \end{cases}$$

and, according to (5.82a), the variation of ϕ in a region of the size of the event horizon radius $H^{-1}(\phi) \sim M_p^{-1}$ should not exceed the order of:

$$\Delta\phi \sim (\partial_i \phi) M_p^{-1} \sim M_p^2 M_p^{-1} = M_p \ll \phi_0, \tag{5.84}$$

so that the ϕ should be relatively homogeneous in typical causal domains.

Furthermore, taking in consideration that we are dealing with a scalar field, so that local anisotropies may appear only from terms $\partial_i \phi$ (and, correspondingly, of the curvature terms to which it couples), we should have that each causal domain in very early spacetime should be locally approximately isotropic. Thus, it should be locally well approximated as a FLRW universe.

Let us then look at the dynamical equations of this universe dominated by a scalar field ϕ. We have the coupled Einstein equations and Euler-Lagrange equations (here, we insert a Planck mass in the Einstein equations to better vizualize the scales in question for our field):

$$H^2 + \frac{k}{a^2} = \frac{8\pi}{3M_p^2}\left(\frac{\dot{\phi}^2}{2} + \frac{(\nabla\phi)^2}{2} + V(\phi)\right), \tag{5.85a}$$

$$\Box\phi = \ddot{\phi} + 3H\dot{\phi} - \frac{1}{a^2}\nabla^2\phi = -V'(\phi), \tag{5.85b}$$

where we once again emphasize that $\Box$ refers to the covariant D'Alembertian while ∇^2 refers to the Laplacian associated with the *static* metric $\tilde{h}_{ij} = a^{-2} h_{ij}$, such that:

$$\Box\phi \equiv g^{ab}\nabla_a\nabla_b\,\phi, \qquad \nabla^2\phi \equiv \tilde{h}^{ij}\tilde{\nabla}_i\tilde{\nabla}_j\phi, \qquad \tilde{\nabla}_i\tilde{h}_{jk} = 0 = \nabla_a g_{bc}. \tag{5.86}$$

We then have that, for a sufficiently uniform field varying in a sufficiently slow manner, more precisely:

$$\begin{cases} \dot{\phi}^2, (\nabla\phi)^2 \ll V(\phi), & (5.87\text{a}) \\ \ddot{\phi}\,, \frac{1}{a^2}\nabla^2\phi \ll V'(\phi), & (5.87\text{b}) \end{cases}$$

the field equations (5.85) can be well approximated by:

$$H^2 + \cancelto{\ll H^2}{\frac{k}{a^2}} = \frac{8\pi}{3M_p^2}V(\phi), \tag{5.88a}$$

$$3H\dot{\phi} = -V'(\phi). \tag{5.88b}$$

It is then not difficult to see that, for an expanding universe ($\dot{a} > 0$) with a not too steep potential slope $V'(\phi)$ near $\phi \approx \phi_0$, the system rapidly evolves to a regime of exponential expansion, in which the curvature term in the LHS of (5.88a) becomes negligible (in terms of effective relative densities, it means the evolution rapidly makes $\Omega_\phi \to 1$, $|\Omega_k| \ll 1$). For reasons that will become apparent below, this is called the *slow-roll regime*, and the conditions (5.87) are called *slow-roll coditions*. In this regime, we have that:

$$\Big(V'(\phi)\Big)^2 = 9H^2\dot{\phi}^2 \simeq \frac{24\pi}{M_p^2}V(\phi)\dot{\phi}^2 \qquad \Rightarrow \qquad \dot{\phi}^2 \simeq \frac{M_p^2}{24\pi}\frac{\Big(V'(\phi)\Big)^2}{V(\phi)}. \tag{5.89}$$

Particularly, for a power-law potential (5.79):

$$\dot{\phi}^2 = \frac{n^2 M_p^2}{24\pi}\frac{V(\phi)}{\phi^2}, \tag{5.90}$$

so that the general restrictions (5.82) that we had previously imposed in our potential (which entailed $\phi \gg M_p$) will then assevere that, in the slow-roll regime:

$$\phi \gg \frac{n}{4\sqrt{3\pi}}M_p \qquad \Leftrightarrow \qquad \tfrac{1}{2}\dot{\phi}^2 \ll V(\phi). \tag{5.91}$$

This condition then reassures the self-consistency of dynamics withing the slow-roll approximation, and (supplemented by $(\nabla\phi)^2 \ll V(\phi)$) it asseveres that the kinetic energy will be much smaller than potential energy, so that the stress tensor $T_{\mu\nu}$ will indeed be dominated by the potential term in the slow-roll regime:

$$T_{\mu\nu} \approx V(\phi) g_{\mu\nu}. \tag{5.92}$$

This means that we shall have precisely the desired equation of state for an inflationary dynamic, namely $p \approx -\rho$, producing a quasiexponential expansion for this patch of the universe.

Of course, one could then ask for how long these conditions hold, whether that is long enough to sustain an inflationary phase at all, and, if so, how much inflation happens while this phase lasts. Well, from the whole set of conditions that we have imposed and derived above, it is not hard to see that the rate of expansion H of the universe will be much larger than the fractional variation rate of ϕ (and therefore of $V(\phi)$), as well as than the fractional variation rate of H itself, so that we do indeed obtain an approximate de Sitter space:

$$\frac{3H\dot{\phi}}{H^2\phi} = -\frac{3M_p^2}{8\pi\phi}\frac{V'(\phi)}{V(\phi)} \approx -\frac{3n}{8\pi}\frac{M_p^2}{\phi^2} \ll 1$$

(where we have once again estimated for a power-law potential with $n \sim \mathcal{O}(1)$). We then have that:

$$\frac{\dot{\phi}}{\phi}H^{-1} = -\frac{n}{8\pi}\left(\frac{M_p}{\phi}\right)^2 \ll 1 \qquad \Leftrightarrow \qquad \frac{\dot{\phi}}{\phi} \ll H. \tag{5.93}$$

Then, taking a time derivative of eq. (5.87a) (where we already neglect the spatial curvature term), we obtain:

$$2H\dot{H} = \frac{8\pi}{3M_p^2}V'(\phi)\dot{\phi} = \frac{8\pi}{3M_p^2}3H\dot{\phi}^2.$$

Thus:

$$\dot{H} = 3\frac{8\pi}{3M_p^2}\left(\cancelto{\ll V(\phi)}{\frac{\dot{\phi}^2}{2}}\right) \ll 3H^2.$$

If we then drop the 3 factor in this magnitude comparison, we obtain simply:

$$\dot{H} \ll H^2, \tag{5.94}$$

which was precisely the slowness condition for the recession of the horizon that we required to obtain a period of quasiexponential expansion! Then, for any time interval $\Delta t \lesssim H/\dot{H}\ \left(\gg H^{-1}\right)$, we should have:

$$a(t) \approx a_0 e^{Ht}, \qquad t \in [t_p, t_p + \Delta t], \tag{5.95}$$

with:

$$H(\phi) = \left[\frac{8\pi V(\phi)}{3M_p^2}\right]^{1/2}. \tag{5.96}$$

In the meanwhile, the field ϕ, governed by equation (5.88b), slowly evolves towards the minimum of its potential; note that this will be just a first order ODE (since we have taken the field to be spatially homogeneous), whose signs are so that ϕ will be driven down the potential curve (with speed proportional to its slope). In this regime, our system is entirely analogous to a particle (with position coordinate ϕ) subject to some viscous friction rolling down a potentiall well with terminal velocity (see Figure 45); for this reason we call this expanding regime *slow-roll inflation*.

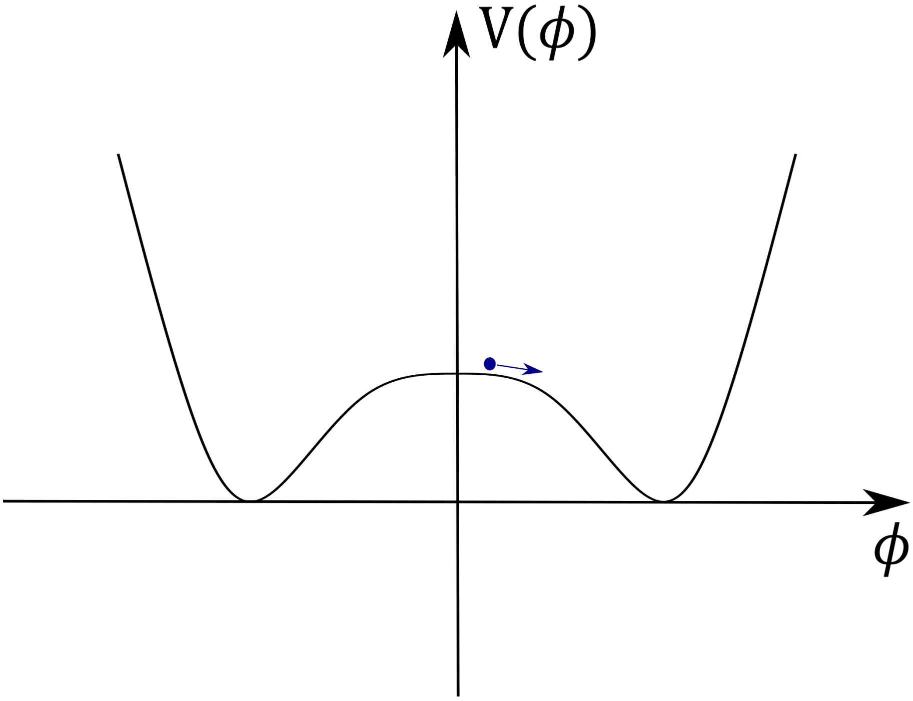

Figure 45 – Field ϕ slowly rolling down its potential well, depicted by a little blue ball which brings out one-dimensional mechanical analog. While the field is near its unstable maximum, it is as though as it is sliding down with terminal velocity in a viscous medium. At later times, when it is considerably far from its maximum, it will perform damped oscillations about its absolute minimum giving up heat to other fields in the universe.
Source: By the author.

5.3.3 Fluctuations of the inflaton field

Having described the *average* behaviour for the evolution of the inflaton ϕ (and the corresponding average evolution of the metric), we shall now take a perturbative approach to describe its fluctuations. As it was described in section 5.2, our approach will consist of splitting our field in an effectively classical contribution ϕ_0 – corresponding to the one that we just described in the previous section, producing an approximately de Sitter background spacetime – and fluctuations $\delta\phi$, which will propagate on this background, and whose own gravitational effects will be neglected on its dynamics. We write this split as:

$$\phi(\mathbf{x}, t) = \phi_0(t) + \delta\phi(\mathbf{x}, t). \tag{5.97}$$

Then, assuming the background solution for $\phi_0(t)$ and $a(t)$ given in the last subsection, we can write the linearized equations for $\delta\phi$ from (5.85b):

$$\frac{\partial^2(\delta\phi)}{\partial t^2} + 3H\frac{\partial(\delta\phi)}{\partial t} - e^{-2Ht}\nabla^2(\delta\phi) = -V'(\phi_0 + \delta\phi) \simeq -V''(\phi_0)\delta\phi. \tag{5.98}$$

Since in this slow-roll regime $V''(\phi_0)$ will be nearly constant, this approximate equation will be simply the one for a free scalar field, with an effective quadratic mass $\mu^2(\phi_0) \equiv V''(\phi_0)$. Well, this is exactly the equation (4.106), that we thoroughly analyzed in section 4.2.4, where μ^2 plays the role of the effective mass M^2:

$$\left[\partial_t^2 + 3H\partial_t - 2e^{-2Ht}\nabla^2 + \mu^2\right]\delta\phi = 0. \tag{5.99}$$

Then, these perturbative variations to the field can be quantized by our basic prescription in chapter 3, and we can employ the same field modes as section 4.2.4, which will correspond to the Bunch-Davies vacuum. However, as μ^2 is dictated by the *average field* ϕ_0, it can have either a positive or a negative sign, depending on the behaviour of its potential; in the slow-regime, it is expected to be slightly negative, as it is decaying near an unstable vacuum (and consistency with the slow-roll conditions imply $\mu^2 \ll H^2$). This will prevent us from straighforwardly transposing the discussion in 4.2.4 to analyze the renormalized expectation value of the fluctuatios $\langle 0|(\delta\phi)^2|0\rangle$ (recall that, in our analysis for the renormalization of the power spectrum, we had pathological IR divergences for negative M^2).

Then, to be able to extract any meaningful information of this fluctuation spectrum, we shall (i) take advantage of the slow-roll mass consistency condition $\mu^2 \ll H^2$ to simplify our field modes in a massless approximation and (ii) argue that, for a space that is not eternally de Sitter, but rather has been inflating for a finite time, it is reasonable to impose an IR cutoff in the spectrum (which shall be implemented at wavelengths many orders of magnitude larger than the hubble radius H^{-1}). The first approximation will imply that $\nu \simeq 3/2$ in equation (4.110). This yields a particularly simple form for the field modes, since the Hankel function $H^{(1)}_{3/2}$ can be put in the form:

$$H^{(1)}_{3/2}(x) = -\sqrt{\frac{2}{\pi x}} e^{-ix} \left(1 + \frac{i}{x}\right). \tag{5.100}$$

Thus we find that the de Sitter adiabatic modes (4.120) will be:

$$h_k(t) = -\frac{iH}{\sqrt{2k^3}} \left(1 + \frac{ik}{H} e^{-Ht}\right) \exp\left(-\frac{ik}{H} e^{-Ht}\right). \tag{5.101}$$

As the universe expands, each of these modes will have its wavelengths exponentially stretched. We then find that, for sufficiently large times, the modes will gradually "exit the horizon" (*i.e.*, reach wavelengths greater than the Hubble radius H^{-1}), which will be given by each mode by:

$$ke^{-Ht} < H \Leftrightarrow t > H^{-1} \ln |k/H|. \tag{5.102}$$

At this point, we see in equation (5.101) that h_k ceases to oscillate and asymptotically 'freezes' in the value:

$$\tilde{h}_k = -\frac{iH}{\sqrt{2k^3}} \tag{5.103}$$

(less of an arbitrary global phase for each mode, which we have not specified). Thus, the dynamical effects of these modes *after inflation* should only manifest when they "reenter" the horizon, and their amplitudes at this point should *only depend on their amplitudes at horizon exit*, which will have happened *during* inflation [U].

If we then analyze the formal expectation value $\langle \phi^2(x) \rangle$ in the (Bunch-Davies) de Sitter vacuum, we obtain the following spectral contributions[V]:

[U] Note that these modes *continue to expand their wavelenths after inflation.* However, after the quasiexponential expansion ceases, the Hubble radius $H^{-1}(t)$ will rapidly recede, allowing reentrance.

[V] The point has been raised by Parker (see *e.g.* (43)) that a physical analysis should consider

$$\langle 0|\phi^2(x)|0\rangle = \frac{1}{(2\pi)^3}\int d^3\mathbf{k}\left(\frac{e^{-2Ht}}{2k} + \frac{H^2}{2k^3}\right), \tag{5.104}$$

which can be expressed in terms of the physical momentum (as measured by a comoving observer) $\mathbf{p} = \mathbf{k}e^{-Ht}$:

$$\langle 0|(\delta\phi)^2(x)|0\rangle = \frac{1}{(2\pi)^3}\int \frac{d^3\mathbf{p}}{p}\left(\frac{1}{2} + \frac{H^2}{2p^2}\right). \tag{5.105}$$

In this form, one can immediately recognizes a 'Minkowski vacuum' type of contribution in the first term (which is only UV-divergent), whereas the second yields an extra inflationary contribution (which is both IR- and UV-divergent)[W]. As we mentioned above, we shall be particularly interested in the very long wavelength behaviour $p = ke^{-Ht} \lesssim H$, for which we can neglect the first term. We then argue that, since physically we do not have an eternally inflating de Sitter space extending all the way past to $t \to -\infty$, but rather a finite inflation phase which cannot be extrapolated past Planck time, it should be reasonable to consider an IR cutoff in the spectrum, restricting our integral to modes with wavelengths within the horizon at $t \sim t_p$ (roughly, with $k = pe^{Ht} > H$). In this case, we obtain the long-wavelength (LW) contribution to the spectrum:

$$\begin{aligned}\langle(\delta\phi)^2\rangle_{LW} &\approx \frac{H^2}{2(2\pi)^3}\int_{LW}\frac{d^3\mathbf{p}}{p^3} \\ &= \frac{H^2}{4\pi^2}\int\limits_{He^{-Ht}}^{H}\frac{dp}{p} = \frac{H^2}{4\pi^2}\int\limits_{H}^{He^{Ht}}\frac{dk}{k} \\ &= \frac{H^2}{4\pi^2}Ht.\end{aligned} \tag{5.106}$$

Note that this linearly increasing time dependence (which appeared from our cutoff when we restricted the spectrum to modes that were *inside* the horizon before inflation)

the *renormalized* value of this spectrum. However, the unsubtracted power spectrum still yields the observed scale-invariant fluctuation spectrum, the main difference being in its associated amplitudes and how they bind the parameters on the inflaton field. Furthermore, the subtracted spectrum is not generally positive-definite, from which difficulties may arise in interpreting it (for a recent discussion in IR divergences and positive definiteness of the spectrum in case of a massless field, see (38)). It is also worth pointing that we are interested in the IR end of the spectrum, whereas the subtraction procedures such as adiabatic subtraction are in principle designed to correct its behaviour in the UV.

[W] As we have previously mentioned, one can cover de Sitter spaces with many coordinate systems (includind static coordinates) and build different vacuum modes associated to different de Sitter Symmetries. For an appropriate choice of modes one could interpret this extra IR-divergent term as being due to particles (see chapter 7 of (4)), with occupation numbers given by $n(\mathbf{p}) = \frac{H^2}{2p^2}$.

can be interpreted as reflecting the fact that, as time passes, more modes exit the horizon (each logarithmic interval of k yielding a similar contribution); of course, it should only be considered for time intervals *during* inflation. Then, precisely from these contributions, one could expect to obtain a (logarithmically) scale-invariant power spectrum of fluctuations for long wavelength modes. Particularly, considering the contribution from a limited spectral integral, say, which exited the horizon during 1 e-folding (*i.e.*, during a time interval $\Delta t \sim H^{-1}$), we obtain:

$$\langle(\delta\phi)^2\rangle_{\Delta k} \approx \frac{H^2}{4\pi^2}. \tag{5.107}$$

5.3.4 Comments on the end of inflation and its physical imprints

Having derived the conditions for the occurrence of an inflationary period, as well as the approximate dynamical equations well within the inflationary phase, we turn our attention to a few quantities that should be of physical significance *after* inflation, and that possibly yield observational consequences today. They are: (i) the total duration of the inflation, (ii) the expansion factor by which the universe inflates in this period, and (iii) the allocation of the energy of the inflaton field ϕ after inflation – both on average and for its fluctuations – and how that influences the subsequent dynamics of the observable universe.

These quantities are relevant to determine whether inflation is a viable candidate to solving the problems in the ΛCDM model in the first place (and if it does not create new problems), as well as if it entails any observational consequences other than those that it was designed to fit.

Let us begin by analyzing the duration of the inflationary phase. To do so, we first consider the evolution of the field well within the slow-roll regime, for which ϕ evolves by a simple first-order equation (5.88b). Then, by substituting the potential (5.79), we end up with a separable equation:

$$\dot{\phi} = \sqrt{\frac{n\lambda_n}{24\pi}} M_p^{3-\frac{n}{2}} \phi^{\frac{n}{2}-1}, \tag{5.108}$$

whose solutions are:

$$\begin{cases} \phi(t) = \phi_0 e^{\sqrt{\frac{\lambda_4}{6\pi}} M_p t}, & n = 4 \qquad \text{(5.109a)} \\ \phi(t) = \left[\phi_0^{\frac{4-n}{2}} - \frac{4-n}{2} \sqrt{\frac{n\lambda_n}{24\pi}} M_p^{3-\frac{n}{2}} t \right]^{\frac{2}{4-n}}, & n \neq 4 \qquad . \qquad \text{(5.109b)} \\ \quad \Rightarrow \; \phi(t) = \phi_0 - \frac{\sqrt{\lambda_2} M_p^2}{2\sqrt{3\pi}} t, & n = 2 \qquad \text{(5.109c)} \end{cases}$$

With those, one can make a rough estimate of when inflation ends by analyzing when there is a significant departure from the slow-roll conditions. Particularly, we can analyze when kinetic energy becomes comparable to potential energy: $V(\phi) \sim \dot{\phi}^2/2$. From equations (5.90) and (5.91), see that this condition should be violated when:

$$\phi \sim \phi_T \equiv \frac{n}{4\sqrt{3\pi}} M_p. \tag{5.110}$$

Then, by inverting equations (5.109), we obtain an estimate for the total duration t_T of inflation as a function of the parameters in our potentials. Particularly, for the quadradic (5.109c) and quartic (5.109a) cases:

$$\begin{cases} t_T \sim \frac{2\sqrt{3\pi}}{\sqrt{\lambda_2} M_p^2} \phi_0, & n = 2 \qquad \text{(5.111a)} \\ t_T \sim \sqrt{\frac{6\pi}{\lambda_4}} M_p^{-1} \ln\left| \frac{\phi_T}{\phi_0} \right|, & n = 4 \qquad \text{(5.111b)} \end{cases}$$

Furthermore, manipulating equations (5.88), it is not difficult to obtain an exact solution $a\big(\phi(t)\big)$ for the slow-roll approximation with this potential:

$$\begin{aligned} \frac{d}{dt} \ln(a) = H &= \frac{8\pi}{3M_p^2} H^{-1} V(\phi) \\ &= -\frac{8\pi}{3M_p^2} 3\dot{\phi} \frac{V(\phi)}{V'(\phi)} \\ &= -\frac{8\pi}{M_p^2} \dot{\phi} \frac{\phi}{n} \\ &= -\frac{4\pi}{nM_p^2} \frac{d\phi^2}{dt}. \end{aligned}$$

Then, we have simply $d(\ln a) = -\frac{4\pi}{nM_p^2} d(\phi^2)$, which yields:

$$a(t) = a_0 e^{\frac{4\pi}{nM_p^2} \left(\phi_0^2 - \phi^2(t) \right)}. \tag{5.112}$$

Of course, for sufficiently small time intervals (for which $\phi_0^2 - \phi^2(t) \approx H(\phi_0)t$), this should be well approximated by (5.95). Still, (5.112) should give us a more accurate estimate for the total inflation factor P as ϕ evolved from $\phi_0 \gg M_p$ to $\phi_T \sim M_p$:

$$P \sim e^{\frac{4\pi}{nM_p^2}\phi_0^2} \sim e^{\frac{4\pi}{n}\left(\frac{\lambda_n}{n}\right)^{-2/n}}. \tag{5.113}$$

Once again, we estimate this factor for quadratic and quartic potentials:

$$\begin{cases} P_T \sim e^{4\pi\frac{M_p^2}{m^2}}, \quad n = 2 & \text{(5.114a)} \\ P_T \sim e^{\frac{\sqrt{2}\pi}{\sqrt{\lambda}}}, \quad n = 4 & \text{(5.114b)} \end{cases}$$

Of course, if we do not have any independent constraints for the potential parameters, little more can be said about the duration of inflation or the inflating factor than the obvious bounds that it should not last up until times where we reach well tested energies, and that it should last long enough to sufficiently dilute spatial curvature and cosmological defects. However, there is a factor that provides us with a more strict estimate of how long inflation should have lasted: the relative density fluctuations in the early universe. As we have seen in the previous section, the average amplitudes of field fluctuations in an exponentially expanding (de Sitter) universe should 'freeze out' when their wavelengths stretch up to H^{-1}, yielding a spectral contribution to fluctuations (5.107):

$$\delta\left<\phi\right> \equiv \sqrt{\left<(\delta\phi)^2\right>_{\Delta k}} \sim \frac{H(\phi)}{2\pi}. \tag{5.115}$$

Then, if we want these field fluctuations to be the ones responsible for the density fluctuations in the CMB, matching an approximately scale-independent spectral amplitude of the order

$$\frac{\delta\rho(k)}{\rho} \sim 10^{-5}, \tag{5.116}$$

we must tune our potential parameters correspondingly. If these fluctuations are indeed produced *during* inflation (due to primordial fluctuations in the inflaton field that later are imprinted in ordinary matter through their interactions) we could make a 'handwaving' estimate of their expected amplitudes as follows[X] :

[X] For a more thorough derivation, see section 7.5 of (4). Also, for an extensive treatment of fluctuation of various types, for many different field species, we refer the reader to chapter 5 of (42).

$$\frac{\delta\rho}{\rho} \sim \frac{\delta V(\phi)}{V(\phi)} = \frac{V'(\phi)\delta\phi}{V(\phi)}. \tag{5.117}$$

Then, using equations (5.115) and (5.96), we obtain the following estimate considering, for simplicity, modes that would have exited the horizon around the end of inflation $\phi \sim \phi_T$:

$$\begin{aligned} \left.\frac{\delta\rho}{\rho}\right|_{k\sim H(\phi_T)} &\sim \frac{V'(\phi_T)}{V(\phi_T)}\frac{H(\phi_T)}{2\pi} \\ &= \frac{n}{\phi_T}\frac{4\sqrt{2}}{4\sqrt{3\pi}M_p}\left[\frac{\lambda_n\phi_T^n}{nM_p^{n-4}}\right]^{\frac{1}{2}} \\ &= 4\sqrt{2}n^{\frac{n-1}{2}}\sqrt{\lambda_n}. \end{aligned} \tag{5.118}$$

For example, estimates in a $\lambda\phi^4$ model (identifying $\lambda = \lambda_4$) yield roughly:

$$\frac{\delta\rho}{\rho} \sim 10\sqrt{\lambda} \qquad \Rightarrow \qquad \lambda \sim 10^{-12}, \tag{5.119}$$

which results in a total inflating factor (5.114b):

$$P_T \sim e^{\frac{\sqrt{2}\pi}{\sqrt{\lambda}}} \sim 10^{10^5}, \tag{5.120}$$

and a total duration (5.111b):

$$t_T \sim \sqrt{\frac{6\pi}{\lambda}}M_p^{-1}\ln\left|\lambda^{-\frac{1}{4}}\right| \sim 10^{-35}s. \tag{5.121}$$

This total time (5.121) hints to us that such a model could be adequate to handle the monopole problem (although it is usually required that t_T be one or two orders of magnitude larger), whereas (5.120) show that it would be comfortably enough to resolve the flatness and horizon problems. Indeed, this astounding inflation factor is significantly larger than it would be required to dilute spatial curvature below the fine-tuning alluded in section 5.1.3, as well as to inflate a region of the Planck size $l_p \sim 10^{-31}m$ up to scales $\delta l \sim 10^{10^5}m$ many orders of magnitude larger than the size of the observable universe $l_U \sim 10^{26}m$, giving observers within the latter enough time to have had causal contact.

Finally, we briefly comment on the subsequent evolution of the universe after the inflationary phase. In principle, if the universe inflated enough to dilute any spatial curvature and monopoles, we also expect ordinary matter and radiation to be brutally diluted. We know, however, that for time scales $t \gtrsim 10^{-13}s$ we had a hot universe essentially dominated by radiation (and, later, with a significant contribution of baryonic matter). To make the two things compatible, we must assume that there will be interactions between the inflaton field and ordinary matter, and that the former will transfer energy to the latter as it decays towards its stable vacuum, in a process that is called *reheating*[Y] ; of course, we need some form of interaction to imprint the primordial fluctuations $\delta\phi$ from the inflaton field in the primordial plasma whose last scattering surface we observe in the CMB today[Z].

Note that, after ϕ decays from its unstable vacuum, it should subsequently oscillate around its absolute minimum (see Figure 45). During these oscillations, interactions with other matter fields should cause it to emit particles, and evolve towards a thermodynamical equilibrium with them. Then, a corresponding upper bound for the reheating temperature can be estimated for an effective number of relativistic degrees of freedom $N^*(T)$ as (4,14):

$$\frac{\pi^2}{30} N^*(T_R)(kT_R^4) \sim V(\phi \sim \phi_T). \tag{5.122}$$

In this case, if we take for instance $N^* \sim 10^2$ in our quartic model, this would entail:

$$kT_R \sim \lambda^{1/4} M_p \sim 10^{15} GeV, \tag{5.123}$$

which is still around the scale of GUT phase transitions. However, the temperature of reheating generally turns out to be orders of manitude lower to that of thermal equlibrium, due to the inneficiency of reheating if we require the interactions of ϕ to other fields in nature to be sufficiently weak (4). In this case, one can transition from an inflationary phase to a usual hot universe described in our standart model, liberated from the fundamental issues of its extrapolation to arbitrarily early times.

[Y] As the term *recombination*, that appears in the cosmology literature to refer to the combination of protons and electrons at the time of matter-radiation decoupling and of the formation of the CMB, this term is somewhat misleading. It is suggestive that the universe was also hot before inflation (as recombination suggests that protons and electrons were combined before decoupling), which would be a hyphothesis with no support on observations.

[Z] For more details on the processes of reheating, see *e.g.* section 7.9 of (4) or section 4.2 of (42).

6 CONCLUSION

Vacuum is complex. If there is a single sentence that captures the message of this dissertation, we believe this should be it. As surprising as it may seem to our intuition this concept turns out to spawn an incredibly rich structure, which may be intimately related to some of the most profound questions that we have about our own universe.

When we switch our fundamental perspective from the notion of particles to that of fields, such that the former comes to be conceived just as an emergent manifestation of the latter, we find that a corresponding notion of vacuum as "a state devoid of particles" cannot make unambiguous sense. Particularly, in quantum theory, different observers can measure very different particle contents in the same field state, depending on their state of motion (even when they are at the same spacetime region). Moreover, even a fixed inertial observer probing the field in a fixed state may observe very different particle contents at different times, giving rise to the phenomenon of particle creation. Nonetheless, all observers converge on their notions of particle occupations at arbitrarily high-energies, corresponding to increasingly localized short wavelength, which allows for a meaningful, although approximate and nonunique, extension of the concept of vacuum in curved spacetimes, which plays a key role in the renormalization of localized quantities, such as field amplitudes and energy densities.

The vacuum energy density, in particular, is found to play significant roles in many contexts. Even classically it may behave in a nontrivial manner, allowing for the description of a finite inflationary phase for the universe. At a quantum level, however, it reveals an even richer scope of possibilities. As long as one can systematically eliminate the divergencies that appear in our description of quantum field theory, quite surprising physical predictions emerge. Even in a description in flat space, where gravity plays no role whatsoever, it was possible to predict and experimentally verify that the electromagnetic vacuum will present negative pressures and induce an attractive force between two conducting plates. In curved spaces, it is found to give rise to a cosmological-constant kind of term, with constant positive energy and negative pressure of the same magnitude, which could conceivably account for the puzzling cosmic component that we now call Dark Energy. Further still, primordial vacuum fluctuations seem like a promising candidate to explain the tiny fluctuations that we observe in the far background sky, and that were ultimately responsible for the formation of the many structures that we observe in the universe today, including the sun, the earth, and the all life that emerged on it.

Of course, there is still much research to be done in the subject before we can extrapolate from appealing theoretical pictures to making strong claims about the workings of the real world. Particularly, on the observational side, increasingly precise and varied measurements of relics from the early universe should allow for significant, and possibly quite surprising, improvements of our understanding of it. Nonetheless, we hope that the present work may serve as a comprehensible introduction to the theoretical window of such a fascinating subject.

REFERENCES

1 BIRRELL, N. D.; DAVIES, P. C. M. **Quantum fields in curved space**. Cambridge: Cambridge University Press, 1982 (Cambridge monographs on mathematical physics).

2 PARKER, L.; TOMS, D. **Quantum field theory in curved spacetime:** quantized fields and gravity. Cambridge: Cambridge University Press, 2009 (Cambridge monographs on mathematical physics).

3 FULLING, S. A. **Aspects of quantum fields theory in curved space-time**. Cambridge: Cambridge University Press, 1989 (London mathematical society student texts).

4 LINDE, A. **Particle physics and inflationary cosmology**. Reading: Harwood Academic Publishers, 1990 (Contemporary concepts in physics).

5 WALD, R.M. **General relativity**. Chicago: The University of Chicago Press, 1984.

6 MANDL, F.; SHAWN, G. **Quantum field theory**. New York: John Wiley & Sons, 1986.

7 LEMOS, N.A. **Mecânica analítica**. São Paulo: Editora Livraria da Física, 2007.

8 WALD, R.M. **Quantum field theory in curved spacetime and black hole thermodynamics**. Chicago: The University of Chicago Press, 1994.

9 WALD, R.M. Particle and energy cost of entanglement of Hawking radiation with the final vacuum state; **Physical Review D**, v. 100, n. 6, 2019. DOI: <https://doi.org/10.1103/PhysRevD.100.065019>.

10 PENROSE R. **Techniques of differential topology in relativity**. Philadelphia: SIAM, 1972.

11 GARCÍA-PARRADO, A.; SENOVILLA, J. M. M. Causal structures and causal boundaries. **Classical and Quantum Gravity** v. 22, n. 9, p. R1–R84 2005. DOI: <https://doi.org/10.1088/0264-9381/22/9/R01>.

12 LIDDLE, A. **An introduction to modern cosmology**. New York: John Wiley & Sons, 2015.

13 RYDEN, B.S. **Introduction to cosmology**. Cambridge: Cambridge University Press, 2018.

14 KOLB, E. W., TURNER, M. S. **The early universe**. Redwood City: Addison-Wesley, 1993.

15 GLASER, L. STEINHAUS, S. Quantum gravity on the computer: impressions of a workshop. **Universe**, v. 5, n. 1, 2019. DOI: <https://doi.org/10.3390/universe5010035>

16 CHAKRABORTY, S. Boundary terms of the Einstein–Hilbert action. *In*: BAGLA J.; ENGINEER, S. (ed.). **Gravity and the quantum:** pedagogical essays on cosmology, astrophysics and quantum gravity. Cham: Springer, 2007. p. 43-59. (Fundamental theories of physics, v. 187). DOI: <https://doi.org/10.1007/978-3-319-51700-1_5>

17 STROCCHI, F. **Symmetry breaking**. Berlin, Heidelberg: Springer, 2008 (Lecture notes in physics, v. 732). DOI: <https://doi.org/10.1007/978-3-540-73593-9>.

18 FRIEDLANDER, F.G.; JOSHI, M. **Introduction to the theory of distributions**. Cambridge: Cambridge University Press, 1982.

19 CHOQUET-BRUHAT, Y.; DeWITT, C.; DILLARD-BLEICK, M. **Analysis, manifolds and physics:** part I: basics. Amsterdam: North Holland, 1982.

20 ARFKEN, G.B.; WEBER, H.J. **Mathematical methods for physicists**. 6th ed. Boston: Elsevier, 2005.

21 GRADSHTEYN, I.S.; RYZHIK, I.M. **Table of integrals, series, and products**. New York: Academic Press, 2007.

22 SLAVYANOV, S.Y.; LAY, W. **Special functions:** a unified theory based on singularities. Oxford: Oxford University Press, 2000.

23 FEYNMAN, R.P. Space-Time approach to non-relativistic quantum mechanics. **Reviews of Modern Physics**, v.20, n.2, p.367, 1948.

24 FEYNMAN, R.P.; HIBBS, A.R., **Quantum mechanics and path integrals**. Minoela: Dover Publications, 2010. (Dover books on physics).

25 HAWKING, S.W.; ELLIS, G.F.R, **The large scale structure of spacetime**. Cambridge: Cambridge University Press, 1973.

26 TOMS, D.J. **The Schwinger action principle and effective action**. Cambridge: Cambridge University Press, 2007.

27 SCHWINGER, J. **Selected papers on quantum electrodynamics**. Mineola: Dover Publications, 1958.

28 SCHWINGER, J.; The theory of quantized fields I. **Physical Review**, v. 82, p.914-927, 1951. DOI: <https://doi.org/10.1103/PhysRev.82.914>

29 SCHWINGER, J., The theory of quantized fields II. **Physical Review**, v.91, p.713-128, 1953. DOI: <https://doi.org/10.1103/PhysRev.91.713>

30 ZAMBIANCO, M.H. **The issue of time in quantum mechanics**. 2021. 126p. Dissertation (Master in Physics) – Instituto de Física Teórica, Universidade Estadual Paulista, São Paulo, 2021. Avaliable from: <http://hdl.handle.net/11449/216577>. Accessible at: 23 Feb. 2022.

31 HUBBLE, E. A relation between distance and radial velocity among extra-galactic nebulae. **Proceedings of the National Academy of Sciences**, v. 15, n. 3, p. 168-173, 1929. DOI: <10.1073/pnas.15.3.168>.

32 EINSTEIN, A. Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie. **Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften** (Berlin), p. 142-152, 1917. Avaliable from: <https://einsteinpapers.press.princeton.edu/vol6-trans/433>. Accessible at: 9 July 2020.

33 FRIEDMAN, A. Über die Krümmung des Raumes. **Zeitschrift für Physik**, v. 10, p. 377–386, 1922. DOI: <https://doi.org/10.1007/BF01332580>.

34 LEMAITRE, A. G. A homogeneous universe of constant mass and increasing radius accounting for the radial velocity of extra-galactic nebulae. **Monthly Notices of the Royal Astronomical Society**, v. 91, n. 5, p. 483–490, 1931. DOI: <https://doi.org/10.1093/mnras/91.5.483>

35 ROBERTSON, H.P. Kinematics and world-structure. **Astrophysical Journal, v. 82, p.284**, 1935. DOI: <https://ui.adsabs.harvard.edu/link_gateway/1935ApJ....82..284R/doi:10.1086/143681>

36 WALKER, A.G. On Milne's theory of world-structure. **Proceedings of the London Mathematical Society**, p. 90-137, 1937. DOI: <https://doi.org/10.1112/plms/s2-42.1.90>

37 FORD, H.L.; PARKER, L. Infrared divergences in a class of Robertson-Walker universes. **Physical Review D**, v. 16, n. 2, 1977. DOI: <https://doi.org/10.1103/PhysRevD.16.245>.

38 ZHANG, Y.; WANG, B.; YE, X. A massless scalar field in Robertson-Walker spacetimes: adiabatic regularization and Green's function. **Chinese Physics C**, v. 44, n. 9, 2020. DOI: <https://iopscience.iop.org/article/10.1088/1674-1137/44/9/095104>

39 ALLEN, B. Vacuum states in de sitter space. **Physical Review D**, v. 32, n. 12, p. 3136-3149, 1985. DOI: <https://journals.aps.org/prd/abstract/10.1103/PhysRevD.32.3136>

40 ALLEN, B.; FOLLACI A. Massless minimally coupled scalar field in de Sitter space. **Physical Review D**, v. 35, n. 12, p. 3771-3778, 1987. DOI: <https://journals.aps.org/prd/abstract/10.1103/PhysRevD.35.3771>.

41 AZUMI, M. *et al.* LiteBIRD: a Satellite for the studies of B-mode polarization and inflation from cosmic background radiation detection. **Journal of Low Temperatures Physics**, v. 194, p. 443–452, 2019. DOI: <https://doi.org/10.1007/s10909-019-02150-5>.

42 WEINBERG, S. **Cosmology**. Oxford: Oxford University Press, 2008.

43 PARKER, L. Amplitude of perturbations from inflation, 2007. Available from: <https://arxiv.org/abs/hep-th/0702216>. Accessible at: 23 Jan. 2021.

44 DI VALENTINO E. *et al.* In the realm of the Hubble tension—a review of solutions. **Classical and Quantum Gravity** v. 38, n. 15. DOI: 10.1088/1361-6382/ac086d

45 HAWKING S.W. Particle creation by black holes. **Communications in Mathematical Physics**, v. 43, n. 3, p. 199-220. DOI: 10.1007/BF02345020

46 UNRUH W.G. Notes on black hole evaporation. **Physical Review D**, v. 14, n. 4, p. 870-892. DOI: <https://doi.org/10.1103/PhysRevD.14.870>

Appendix

APPENDIX A – DISTRIBUTIONS

The subject of distributions is one hard to ignore in physics, and yet it is seldom given a proper treatment in the exposition of topics for which it is relevant (ranging from point-charges in electrostatics, going through bras and kets in ordinary quantum mechanics, and up to its ubiquitous presence in field theory). The present exposition of the topic, far from exhaustive or fully rigorous, aims at laying a few basic definitions and providing a clear picture for its applications in the scope of this dissertation. For a longer but straightforward and physically-oriented exposition, we recommend (18); or, for a more general and rigorous covering of the topic, see (19).

- **Distributions as linear functionals.**

The subject of distributions is one of linear algebra. When one is handling an ordinary finite-dimensional vector space $\mathbb{V}$, one fundamental concept is that of linear operators acting on $\mathbb{V}$ to produce scalars. Those operators, in their turn, form another linear space, the *dual space* $\mathbb{V}^*$. $\mathbb{V}^*$ can be easily shown to be isomorphic to $\mathbb{V}$, although there is no natural identification between them.[A]

In this case, let $n = \dim(\mathbb{V}) = \dim(\mathbb{V}^*)$ and consider an arbitrary pair of a vector $v \in \mathbb{V}$ and a dual vector $\sigma \in \mathbb{V}^*$. Given a basis $\{\mathbb{e}_i\}$ of $\mathbb{V}$ and its dual basis $\{\mathbb{e}_i^*\}$ in $\mathbb{V}^*$ ($\mathbb{e}_i^*(\mathbb{e}_j) = \delta_{ij}$), we can write $\sigma(v)$ in terms of their components:

$$\sigma(v) = \sum_{i=1}^{n} \sigma_i v^i. \tag{A.1}$$

However, in the case of infinite-dimensional vector spaces, such as many function spaces, it is no longer generally true that $\mathbb{V}$ and $\mathbb{V}^*$ are isomorphic. In fact $\mathbb{V}^*$ is generally bigger than $\mathbb{V}$: that means one can generally associate to each vector $v \in \mathbb{V}$ a dual vector $u \in \mathbb{V}^*$, but the converse is not always true. Specifically, for function spaces (usually, for "well-behaved" functions defined on some open set of a sufficiently smooth manifold $\mathcal{O} \subset \mathcal{M}$), these dual vectors are linear *functionals*, and they are called *distributions*.

We shall generally denote the function space under consideration $\mathcal{F}$, whose domain $\mathcal{O}$ is assumed to have a measure μ[B], and the space of distributions that act on it $\mathcal{D}$ ($= \mathcal{F}^*$); then, the action of a distribution on a function will generally take the form of a Lebesgue integral. A case of particular interest is when $\mathcal{F}$ is a space of smooth functions on open

[A] On the other hand, the dual of $\mathbb{V}^*$, $\mathbb{V}^{**}$, can be naturally identified with $\mathbb{V}$: given $V \in \mathbb{V}^{**}$ one associates it with the unique vector $v \in \mathbb{V}$ that satisfies $V(\sigma) = \sigma(v)$, $\forall \sigma \in \mathbb{V}^*$.

[B] For a metric manifold $(\mathcal{M}, g_{ab})$, this just amounts to familiar integrals in $\mathbb{R}^d$: $\int_{\mathcal{M}} d\mu_g(x) f(x) = \int_{\mathbb{R}^d} d^d x |g(x)|^{\frac{1}{2}} f(x)$.

intervals $U \subset \mathbb{R}^d$, $\mathcal{C}^\infty(U)$ (the case of smooth manifolds can be mapped in (a countable sum of) this one[C]);0 for a brute-force guarantee that one will not have to worry with boundary terms, one often resorts to the subset of these functions with compact support $\mathcal{C}_0^\infty(U)$. Throughout this Appendix, to avoid being cumbersome with technical remarks, we shall always assume our functions to be sufficiently well-behaved so that our assertions make sense (for example, that one may take derivatives to a desired order, that certain integrals converge and that there will be no contributions from boundary terms).

Analogously to (A.1), when we have a distribution $\sigma \in \mathcal{F}^*$ *that can be identified with a locally integrable function* $\tilde{\sigma} \in \mathcal{F}$, we may write its action on any function $f \in \mathcal{F}$ as:

$$\sigma[f] = \int d\mu(x)\tilde{\sigma}(x)f(x). \tag{A.2}$$

However, that is not always the case. Consider the linear functional δ_x, $x \in \mathcal{O}$, whose action upon a function $f \in \mathcal{F}$ produces the value of f in x, that is: $\delta_x[f] \equiv f(x)$. Clearly, there is no locally integrable function $\tilde{\delta}_x$ that satisfies this property in the continuum[D]. However, we would still like to represent this functional in that form:

$$\delta_x[f] = f(x) = \int_{\mathcal{O}} d\mu(y)f(y)\delta(x,y) = \int_U d^dyf(y)\delta(x-y) = \int_U d^dyf(y-x)\delta(y), \tag{A.3}$$

so that one often speaks of the “Dirac delta fuction” $\delta(x)$, as though one was actually integrating a function in (A.3) – or, extrapolating further, one speaks of the delta function *outside an integration sign*, where it makes much less formal sense. Our aim here is not to be pedantic about definitions; on the contrary, *we are interested in potentializing their practical use without incurring in any operational pitfalls.* It is usually harmless, and often quite useful, to make informal manipulations with distributions as though they were functions (as it is to “pass dx multiplying” in a differential equation, or use “the wave-function of a particle of momentum k”: $\psi_k(x) = (2\pi)^{-\frac{1}{2}}e^{ikx}$). However, there are situations where such approaches break and produce implausible or paradoxical results, and that often happens because one failed to appreciate that he/she is not dealing with a function, but rather with a distribution, and that more care to a particular operation was needed (just like one may be led to the absurd conclusion that $1 = 2$ by carelessly performing algebraic manipulations involving a division by 0). A few sensible/troublesome manipulations with distributions include trying to evaluate objects like $\langle x|x\rangle$, $\int \delta^2(x)f(x)dx$, which would be perfectly natural if one was operating ordinary functions or vectors.

In the context of ordinary nonrelativistic quantum mechanics, one finds that plane waves, the eigenfunctions of the momentum operator, are not square-integrable in $\mathbb{R}^n$, and,

[C] See appendix B of (5) for details of integration on manifolds.

[D] For a slightly longer discussion of this point, see chapter 1 of (18) as such, they cannot properly represent state vectors in the Hilbert space $\mathcal{H} = \mathcal{E} \subset \mathcal{L}^2(\mathbb{R}^n)$ (we take a subset $\mathcal{E}$, not the entire $\mathcal{L}^2$, to ensure we are restricted to sufficiently well-behaved functions). The case is even worse for the eigenvectors of the position operator[E], which are not even functions to start with. Nevertheless, all *actual* wave-functions $\psi \in \mathcal{E}$ have well-defined Fourier transforms, as well as (obviously) well-defined values at each point. Thus, although position and momentum eigenstates $|x\rangle : X\,|x\rangle = x\,|x\rangle$ and $|p\rangle : P\,|p\rangle = p\,|p\rangle$ do not make rigorous sense as state vectors belonging to the Hilbert space $\mathcal{H}$, it is perfectly sensible to evaluate them as distributions: $\forall\,|\psi\rangle \in \mathcal{H}$,

$$\psi(x) = \langle x|\psi\rangle\,, \quad \langle x|X|\psi\rangle = x\psi(x), \quad |\psi\rangle = \int dx\,|x\rangle\,\langle x|\psi\rangle = \int dx\,|x\rangle\,\psi(x), \tag{A.4}$$

$$\bar{\psi}(p) = \langle p|\psi\rangle\,, \quad \langle p|P|\psi\rangle = p\bar{\psi}(p), \quad |\psi\rangle = \int dp\,|p\rangle\,\langle p|\psi\rangle = \int dp\,|p\rangle\,\bar{\psi}(p). \tag{A.5}$$

We also have the famous orthornormality relations, which can be rigorously stated in the distributional sense: $\langle x|x'\rangle = \delta(x-x')$, $\langle p|p'\rangle = \delta(p-p')$, $\langle x|p\rangle = (2\pi)^{-1/2}e^{\pm ipx}$, and which allow one to evaluate any scalar products $\langle\phi|\psi\rangle$ or operator transformations $\langle\phi|A|\psi\rangle$ in terms of the $\{|x\rangle\}$ or $\{|p\rangle\}$ bases.

- **Field operators as distributions and applications to 2-point functions:**

Similarly, in the context of QFT, it occurs that the fundamental observables (namely, field configurations or canonical momenta configurations) do not have proper eigenstates in their respective Hilbert Spaces. Nonetheless, one often just writes:

$$|\phi'\rangle : \phi(x)\,|\phi'\rangle = \phi'(x)\,|\phi'\rangle\,, \quad \forall x \in \Sigma, \tag{A.6}$$

$$|\pi'\rangle : \pi(x)\,|\pi'\rangle = \pi'(x)\,|\pi'\rangle\,, \quad \forall x \in \Sigma, \tag{A.7}$$

where we used primes $'$ to distinguish between the field (momentum) operator $\phi(x)$ $(\pi(x))$ – evaluated at any event x belonging to a certain Cauchy surface[F] Σ – and its eigenvalue $\phi'(x)$ $(\pi'(x))$.

To be more precise, the quantized field observable $\phi(x)$ is actually an *operator-valued distribution* (such that $\langle\Psi_1|\phi(x)|\Psi_2\rangle$ is an ordinary (number-valued) distribution, $\forall\,|\Psi_1\rangle\,, |\Psi_2\rangle \in \mathcal{H}$). (Further considerations an implications of that point can be found in chapter 3 of (3))

[E] Note that these are not vectors belonging to $\mathcal{H}$, but rather to a larger space: $\mathcal{H}^{**}$ (the dual of the dual of $\mathcal{H}$).

[F] Note that we cannot have a *single* eingenstate of the field operator $\phi(x)$ *in all spacetime*, just like we cannot have a wave-function perfectly localized *at all times*: this would amount to having a sharp (classical) trajectory for ϕ. Determining it (*or* its momentum) at an entire simultaneity surface Σ corresponds to obtaining its maximal information in a quantum description.

There are some basic operations that one may perform with distributions. Sum and mulplication by scalars are the most elementary ones, stemming directly from the vector space structure of $\mathcal{D}$. Further uselful operations that one may perform with distributions are *tensor products*, *convolutions* and *taking derivatives*. All of them are well-defined and quite intuitive to handle operationally by recurring to their function correspondents; we refer the reader to (18) for a more detailed definition and examples of each operation (see, respectively, chapters 2, 4 and 5).

Therefore, we have that bilinear objects such as $\phi(x)\phi(x')$, or any combination of them that involves derivatives (such as the terms that appear in the two-point stress tensor $T_{\mu\nu}(x,x')$), are generally well-defined as distributions. They will often show a singular behaviour when one attempts to evaluate them as $x \to x'$, whereas they are usually regular for $x \neq x'$. Such a behaviour is not at all surprising when we think of the paradigmatic example of Dirac deltas: If we define a test function $F \in \mathcal{F} \otimes \mathcal{F}$, it is perfectly sensible to evaluate the 'double delta' distribution $\Delta(x,x') \equiv \delta_x \otimes \delta_{x'} \in \mathcal{D} \otimes \mathcal{D}$:

$$\begin{aligned}\Delta(x,x')[F] &\equiv \iint d\mu(y)d\mu(y')F(y,y')\delta(x,y)\delta(x',y') && \text{(A.8)}\\ &= F(x,x'). && \text{(A.9)}\end{aligned}$$

It is also perfectly sensible to evaluate the convolution of two Dirac deltas, $(\delta * \delta)_x$, in a test function $f \in \mathcal{F}$:

$$\begin{aligned}(\delta * \delta)_x[f] &\equiv \int dy f(x-y)(\delta * \delta)(y)\\ &\equiv \iint dydz f(x-y)\delta(y-z)\delta(z)\\ &= \int dy f(x-y)\delta(y)\\ &= f(x). && \text{(A.10)}\end{aligned}$$

So one immediately finds that $(\delta * \delta)_x = \delta_x$ (of course, this last equality is not true for *any* distribution σ, *i.e.* $\sigma_x \not\equiv (\sigma * \sigma)_x$). Now, *the product of 2 distributions is not generally defined*. Although for any two functions, g, and u, one could evaluate their action in a test function f like:

$$(gu)[f] = \int dx\, g(x)u(x)f(x), \qquad \text{(A.11)}$$

the same is *not true* for two distributions σ and θ, for which it generally makes no sense to evaluate:

$$(\sigma\theta)[f] = \int \sigma(x)\theta(x)f(x), \tag{A.12}$$

obvious exceptions being the case where one of these distributions can be identified with a function (or when $\mathrm{supp}(\sigma) \cap \mathrm{supp}(\theta) = \emptyset$, for which we could trivially define $\sigma\theta = 0$). Particularly, there is no direct way to make sense of an object like $\delta^2(x)$.

Then, it should not be surprising that the attempt to directly evaluate expected values such as $\langle\Psi|\phi^2(x)|\Psi\rangle$ does not make direct sense, and generally yields divergent results. Much more appaling is the fact that, for a quite large variety of field theories, these divergencies can actually be systematically handled and subtracted to yield meaningful finite physical results (although these procedures often require very sophisticated techniques and cumbersome calculations, and they are not generally free from ambiguities).

- **Discontinuities and singularities; principal value of distributions:**

In field theory, it is also not unusual that one must handle distributions involving integrals that go directly through singularities in their integrands. In such cases, there is a variety of ways through which one may obtain a meaningful value of the integration, giving rise to ambiguities to such singular distributions. Among the many ways to *define* the action of singular distributions, one conventional one is their so-called *principal value*. Ultimately, evaluating the principal value is one convenient way to cancel out infinities and obtain meaningful finite results; here, we shall give just a superficial glimpse in the subject, applying it to simple distributions that appear in this dissertation (again, we refer the reader to chapter 2 of (18) for a more rigurous and thorough exposition of the subject)

A case of particular interest to us will be calculating the principal value of integrals around order 1 poles. Thus, for a start, let us consider a functional $\sigma \in \mathcal{D}(\mathcal{C}_0^\infty)$ defined through the function $1/x$, which has a singularity at $x = 0$. How, then, should we interpret its action on a test function $f \in \mathcal{C}_0^\infty$, $\sigma[f]$? Well, since $\frac{1}{x}$ can be written as $\frac{d}{dx}\ln|x|$ for $x \neq 0$, *one particular* way to evaluate it is:

$$\begin{aligned}
\sigma[f] &= \int_{-\infty}^{\infty} dx \frac{f(x)}{x} = -\int_{-\infty}^{\infty} dx f'(x)\ln|x| \\
&\equiv -\lim_{\epsilon\to 0^+}\left[\int_{-\infty}^{-\epsilon} dx f'(x)\ln(-x) + \int_{\epsilon}^{\infty} dx f'(x)\ln(x)\right] \\
&= \lim_{\epsilon\to 0^+}\left[\cancelto{\mathcal{O}(\epsilon\ln\epsilon)\to 0}{\Big(f(\epsilon) - f(-\epsilon)\Big)}\ln(\epsilon) + \int_{-\infty}^{-\epsilon} dx \frac{f(x)}{x} + \int_{\epsilon}^{\infty} dx \frac{f(x)}{x}\right] \\
&= \lim_{\epsilon\to 0^+}\left[\int_{-\infty}^{-\epsilon} dx \frac{f(x)}{x} + \int_{\epsilon}^{\infty} dx \frac{f(x)}{x}\right]
\end{aligned}$$

$$\equiv \mathcal{P}\left(\int_{-\infty}^{\infty} dx \frac{f(x)}{x}\right), \tag{A.13}$$

which is how we *define* the principal value of $\frac{1}{x}$, $\mathcal{P}\left(\frac{1}{x}\right)$.

Note that (i) the second equality is a particular (arbitrary) way to take the limit in the domain around $x = 0$ (the results could be different if we approached 0 at different rates from the left and the right) and (ii) since the logarithm is also singular at $x = 0$ (although it is the derivative of a piecewise continuous function) this integration by parts is also not free of ambiguities. For instance, we could have defined:

$$\begin{cases} \frac{1}{x} = \frac{d}{dx}\ln(-x), & x < 0 \\ \frac{1}{x} = \frac{d}{dx}\ln(x) + a, & x > 0 \end{cases},$$

so that the same procedure would yield:

$$\begin{aligned}
\sigma[f] = \int_{-\infty}^{\infty} dx \frac{f(x)}{x} &= -\int_{-\infty}^{0} dx f'(x)\ln(-x) + \int_{0}^{\infty} dx f'(x)\Big(\ln x + a\Big) \\
&\equiv -\lim_{\epsilon\to 0^+}\left[\int_{-\infty}^{-\epsilon} dx f'(x)\ln(-x) + \int_{\epsilon}^{\infty} dx f'(x)\ln(x) + a\int_{\epsilon}^{\infty} dx f'(x)\right] \\
&= \lim_{\epsilon\to 0^+}\left[\int_{-\infty}^{-\epsilon} dx \frac{f(x)}{x} + \int_{\epsilon}^{\infty} dx \frac{f(x)}{x}\right] + af(0) \\
&\equiv \mathcal{P}\left(\int_{-\infty}^{\infty} dx \frac{f(x)}{x}\right) + af(0).
\end{aligned} \tag{A.14}$$

This gives us the distribution $\sigma(x) \equiv \mathcal{P}(\frac{1}{x}) + a\delta(x)$. Since taking the value $a = 0$ is just an arbitrary choice as any other value, so will be the result of the distribution that we try to associate to $1/x$, in regards to its singular region. The point here is that the behaviour of the function in its nonsingular region does not uniquely determine a distribution, and additional information regarding its singular region may be required to define it. For a order 1 pole, the particular way of defining its principal value is adding up the divergent contributions aroud 0 symmetrically, so that they cancel out.

Another interesting application of considering a $1/x$ distribution as the derivative of $ln(x)$ emerges when we consider functions in the complex plane. Let z be a complex number, $z = x + iy = |z|e^{i\arg(z)}$, so that its logarithm is defined as:

$$\ln(z) = \ln|z| + i\arg(z). \tag{A.15}$$

Then, if we take the limit $y \to 0^+$, we have for all finite x that:

$$\lim_{y\to 0^+} \ln(z) = \ln|x| + i\pi(1-\Theta(x)) \qquad \Rightarrow \qquad \frac{d}{dx}\Big(\lim_{y\to 0^+}\ln(z)\Big) = \frac{1}{x} - i\pi\delta(x). \tag{A.16}$$

Similarly, for negative values of y:

$$\lim_{y\to 0^-} \ln(z) = \ln|x| - i\pi(1-\Theta(x)) \qquad \Rightarrow \qquad \frac{d}{dx}\Big(\lim_{y\to 0^-}\ln(z)\Big) = \frac{1}{x} + i\pi\delta(x). \tag{A.17}$$

We can then use the complex identity $z^{-1} = \frac{d}{dz}\ln(z)$ to obtain the following distribution in the reals:

$$\sigma(x) = \lim_{\epsilon\to 0^+} \frac{1}{x \pm i\epsilon} \equiv \mathcal{P}\left(\frac{1}{x}\right) \mp i\pi\delta(x). \tag{A.18}$$

With this identity, one can analyze (particular values of) integrals along the real axis, by displacing their poles infinitesimally in the complex plane (either above or below the axis, depending on the application at hand).

APPENDIX B – SOME GEOMETRICAL DERIVATIONS

Most texts in QFTCS already assume the reader to be familiar, to a fair extent, with both QFT in Minkowski spacetime and General Relativity. In this work, while we do provide a full introductory chapter to QFT, we shall not present a thorough and comprehensive introduction to GR (for that, we refer the reader to the excellent textbook of R. Wald (5), where this author personally learned the subject; alternatively, see (25)). Still, the need was felt to provide an appendix discussing some fundamentals and covering more specific geometric derivations. It should serve both to lay the basic definitions and notations, and to explicitly develop some useful tools for our discussion in the main text, avoiding gaps in our derivations. Besides defining the fundamental geometrical objects used in the formulation of the theory, such as curvature and covariant derivatives, the topics in this appendix include an introduction to the computations of variations in respect to the metric, as well as a few useful geometrical structures, like Lie derivatives, Killing fields, and conformal transformations.

B.1 Fundamental building blocks of GR

The theory of General Relavity, whose original formulation was culminated in Einstein's work, succeeded to incorporate two very simple founding physical principles[A] in a geometrical formalism for spacetime. In this formulation, spacetime came to be conceived as a curved, pseudo-Riemannian manifold, whose dynamics are governed by the matter propagating in it. Of course, one can trace the roots of this theory back to the simpler geometrical formulation of special relativity, built in Minkowski spacetime $(\mathbb{R}^4, \eta_{ab})$, from which one can find a generalization in curved spacetimes $(\mathcal{M}, g_{ab})$, suitable to general relativity.

However, unlike Minkowski spacetime (or even prerelativistic Galilean spacetime), whose affine structure allows for fairly simple geometrical constructions and manipulations with little more than linear algebra and calculus tools, in curved spaces one requires a quite more sophisticated paraphernalia from differential geometry to carry various relevant calculations.

To start with, one can no longer use a single vector space structure to define vectors v^a, dual vectors (also called *covectors*) ω_a and general tensors $T^{abc\dots}_{def\dots}$ *in the entire spacetime*. Instead, one must build *tangent spaces to each event* $x \in \mathcal{M}$, $\mathbb{V}_x$, and work with tangent vectors $v^a(x)$, dual vectors (also called *cotangent vectors*) $\omega_a(x)$ and general tensors $T^{abc\dots}_{def\dots}(x)$, with no natural identification between $\mathbb{V}_x$ and $\mathbb{V}_{x'}$ for two distinct events

[A] Namely, the local invariance of the speed of light and the equivalence principle.

$x \neq x'$.[B] This, on its turn, prevents one from having a *unique* geometrical notion of derivatives for tensor fields, which must be specified through further physical postulates. First, one usually requires that it acts symmetrically on scalars, that is:

$$\nabla_a \nabla_b \phi - \nabla_b \nabla_a \phi = 0. \tag{B.1}$$

Since generally one could have that:

$$\nabla_a \nabla_b \phi - \nabla_b \nabla_a \phi = T_{ab} \phi, \tag{B.2}$$

being $T_{ab} = -T_{ba}$ the so-called the *torsion tensor*, this is referred to as the *null torsion* condition (or, equivalently, one says that $(\mathcal{M}, g_{ab})$ is a *torsion-free space*).

Generally, two derivative operators ∇_a and $\tilde{\nabla}_a$ may differ in the following way:

$$\tilde{\nabla}_b v^a = \nabla_b v^a + C^a_{\ bc} v^c, \tag{B.3}$$

where $C^a_{\ bc}$ is called a *connection*. For torsion free spaces, it will be symmetrical in the lower indices: $C^a_{\ bc} = C^a_{\ cb}$.

Thus, the fundamental objects in GR are the metric field g_{ab} and a preferred derivative operator ∇_a (or, equivalently, a preferred connection). In the standard formulation of GR, motivated by the equivalence principle, one requires the physical derivative operator – the so-called covariant derivative ∇_a – to be that with respect to which local variations of the metric vanish:

$$\nabla_c \, g_{ab} = 0. \tag{B.4}$$

That is not to say the metric is spacetime homogeneous, but rather that a fundamental notion of a locally nonvarying quantity is defined in closed proximity with it[C].

[B] **Here, we are using the *abstract index notation* (see List of Symbols?); throughout this work, we often switch between concrete (greek) indices and abstract ones (latin, from *a* to *h*), using the former more often in the context of quantum field theory and the latter in 'purely geometrical/relativistic' contexts, maintaining similarities with the literature**.

[C] In section B.3, when we define Lie derivatives and construct the notion of continuous isometry groups, we shall ascribe a clearer meaning to the notions of variations of the metric along spacetime from a purely geometrical perspective.

Although the formulation so far has been carried in a coordinate-free way, one must often adopt a coordinate system to carry calculations. In practice, one must be able to compute covariant derivatives in terms of the metric components $g_{\mu\nu}$ and ordinary coordinate derivatives ∂_μ. These can be calculated as a particular instance of (B.3), using the Christoffel Symbols $\Gamma^a{}_{bc}$:

$$\nabla_\mu v^\nu = \partial_\mu v^\nu + \Gamma^\nu{}_{\mu\alpha} v^\alpha, \tag{B.5}$$

where $\Gamma^\alpha{}_{\mu\nu}$ can be calculated as a function of metric components as:

$$\Gamma^\alpha{}_{\mu\nu} = \tfrac{1}{2} g^{\alpha\beta}\Big[\partial_\mu g_{\nu\beta} + \partial_\nu g_{\mu\beta} - \partial_\beta g_{\mu\nu}\Big]. \tag{B.6}$$

For practical computations, it is very useful to write a contraction of these symbols in termos of *the determinant* of the matrix of metric components, g, which reads:

$$\Gamma^\alpha{}_{\alpha\mu} = \tfrac{1}{2} g^{\alpha\beta}\partial_\mu g_{\alpha\beta} = \tfrac{1}{2} g^{-1}\partial_\mu g = \partial_\mu \ln|g|^{\frac{1}{2}} = |g|^{-\frac{1}{2}}\partial_\mu |g|^{\frac{1}{2}}. \tag{B.7}$$

The most useful application of this formula for us will be in computing the D'Alembertian for a scalar field. Note that, for a scalar field, while first derivatives will be simply given by partial derivatives, second derivatives will involve Christoffel symbols. We can then write:

$$\begin{aligned}\Box\phi \equiv g^{\mu\nu}\nabla_\mu\nabla_\nu\phi &= \nabla_\mu\Big(g^{\mu\nu}\partial_\nu\phi\Big)\\ &= \partial_\mu\Big(g^{\mu\nu}\partial_\nu\phi\Big) + \Gamma^\mu{}_{\mu\alpha} g^{\alpha\nu}\partial_\nu\phi\\ &= \partial_\mu\Big(g^{\mu\nu}\partial_\nu\phi\Big) + g^{\alpha\nu}\partial_\nu\phi\Big(|g|^{-\frac{1}{2}}\partial_\alpha|g|^{\frac{1}{2}}\Big)\\ &= |g|^{-\frac{1}{2}}\partial_\mu\Big(|g|^{\frac{1}{2}} g^{\mu\nu}\partial_\nu\phi\Big)\end{aligned} \tag{B.8}$$

Now, we would like to construct an intrinsic notion of curvature for our spacetime, defined uniquely by the metric g_{ab}. Ultimately, this curvature will refer to *derivatives* of the metric (up to second order). This may sound very strange at this point, since we postulated the covariant derivative of the metric to be identically null in (B.4). It happens that this equation precisely defines how the physical notion of derivatives depend on the metric. Thus, we define the notion of curvature indirectly, through covariant derivatives. To start with, we define the Riemann curvature tensor[D] by:

[D] Conventions may somewhat vary in the literature, due both to the choice of metric signature and conventions on different indices.

$$\nabla_a \nabla_b \, \omega_c - \nabla_b \nabla_a \, \omega_c = -R_{abc}{}^d \omega_d, \tag{B.9}$$

so that, indirectly, one can uniquely associate a curvature tensor to a given metric, by equations (B.4) and (B.9).

One can also cumpute the curvature components in terms of Christoffel Symbols (that is, in terms of partial derivatives of the metric components) as:

$$R_{\mu\nu\alpha}{}^{\beta} = 2\partial_{[\mu} \, \Gamma^{\beta}{}_{\nu]\alpha} + 2\Gamma^{\lambda}{}_{\alpha[\nu} \, \Gamma^{\beta}{}_{\mu]\lambda}. \tag{B.10}$$

It is also worth listing here a few of the symmetries and identities obeyed by the Riemann tensor. They can each be worked with some algebraic effort, and we shall state them without proofs. First we have three independent antissymmetry properties:

$$1. \; R_{abcd} = -R_{bacd}, \tag{B.11a}$$

$$2. \; R_{abcd} = -R_{abdc}, \tag{B.11b}$$

$$3. \; R_{[abc]d} = 0 \tag{B.11c}$$

(where 2 makes explicit use that ∇_a is the covariant derivative in $(\mathcal{M}, g_{ab})$). Together, they imply the following symmetry:

$$R_{abcd} = R_{cdab}. \tag{B.11d}$$

Finally, we state the Bianchi identity:

$$\nabla_{[a} R_{bc]de} = 0. \tag{B.11e}$$

Contractions of the Riemann tensor play a key role in GR. Because of all its antissymetry properties, there is only one independent rank $(0, 2)$ contraction, which will be the Ricci tensor. We then define the Ricci curvature R_{ab} and the curvature scalar as R as:

$$R_{ac} \equiv R_{abc}{}^b, \tag{B.12}$$

$$R \equiv g^{ac} R_{ac}. \tag{B.13}$$

Note that in virtue of eq (B.11d), $R_{ab} = R_{ba}$. An important combination of R_{ab} and R is the so-called *Einstein tensor*, defined as:

$$G_{ab} = R_{ab} - \tfrac{1}{2} R g_{ab}. \tag{B.14}$$

It is this tensor that will appear at the left side of Einstein equations. Finally, we note that the Bianchi identity will imply that G_{ab} is covariantly conserved:

$$\nabla_a G^{ab} = 0. \tag{B.15}$$

B.2 Variations with respect to the metric

As we have seen thus far, the construction of geometrical quantities from the metric is often very indirect, being determined by the particular way that it fixates the covariant derivative in our spacetime. Thus, it is useful to compile a few results for the systematic computation of variations of these quantities as we vary the metric. Our main interest in computing these variations will be to derive functional derivatives of curvature tensors in the context of field theory. However, another very important and immediate application of such results lies in obtaining perturbative solutions for Einsteins' equations near some known solution, so we make our derivations directly in this latter context. Once we have the desired results at hand, we directly interpret them in terms of infinitesimal variations.

We start by considering a dynamic equation for a generic field variable g (concretely, in our context of interest, this will be Einstein's Equations for the metric), which can be put in the form:

$$\mathcal{E}(g) = 0, \tag{B.16}$$

where $\mathcal{E}$ is some local differential functional of g.

Let us supose now that we know an exact solution ${}^0g$ of (B.16), and that we are considering a situation in which the deviation from a certain solution of interest, g, with respect to ${}^0g$ is small. In fact, to express this assertion in a mathematically meaningful way, we assume the existence of a (1-parameter) family of exact solutions of (B.16), $g(\lambda)$, i.e.

$$\mathcal{E}\Big(g(\lambda)\Big) = 0, \qquad \forall \lambda \in I \tag{B.17}$$

(where the domain $I \subset \mathbb{R}$ contains an open interval around 0), such that:

$$\begin{cases} (i)\, g(\lambda) \text{ is a differentiable function of } \lambda & \text{(B.18)} \\ (ii)\, g(0) = {}^0g & \text{(B.19)} \end{cases} .$$

One may then think of λ as a parameter that quantifies the deviation of some (unknown) exact solution $g(\lambda)$ to our known solution ${}^0g$. This way, for arbitrarily small values of λ, we obtain solutions that will be arbitrarily close to ${}^0g$. However, since (B.16) may be too difficult **(or impossible)** to solve exactly, we may then obtain an approximate solution by noting that:

$$\frac{d}{d\lambda}\Big[\mathcal{E}\Big(g(\lambda)\Big)\Big] = 0 \tag{B.20}$$

(since (B.17) is valid for all $\lambda \in I$). Particularly, this equality must hold for $\lambda = 0$. If we then define a perturbation in our field as $\delta g \equiv \lambda\gamma$, that is:

$$\gamma \equiv \left.\frac{dg(\lambda)}{d\lambda}\right|_{\lambda=0}, \tag{B.21}$$

then, by Leibniz's rule, it is easy to see that (B.20) will give us a linear equation for γ, in the form:

$$\mathscr{L}(\gamma) = 0, \tag{B.22}$$

where $\mathscr{L}$ is a linear local differentiable operator. One then calls (B.22) the 'linearization of (B.16) around ${}^0g$'.

Then, our case of interest will be when g represents the metric field g_{ab} and $\mathcal{E}(g_{ab}) = G_{ab}$, such that (B.16) will represent the exact Einstein Equations in the vacuum[E] (which can actually be written simply as $R_{ab} = 0$). In this case, let us explicitly derive the *linearized* equations (B.22) for the perturbation in the metric.

To achieve that, we must calculate the Ricci tensor $R_{ab}(\lambda)$ associated with the metric $g_{ab}(\lambda)$ in a useful expression. More specifically, we want an expression for it in terms of the background metric ${}^0g_{ab}$ and with explicit algebraic functions of λ, so that we may clearly take derivatives with respect to it. The challenge in doing that lies in the fact that the curvature is only indirectly defined in respect to the metric, through covariant derivatives (B.9), which are bound to obey eq (B.4). We then begin by noting that the covariant derivatives ${}^\lambda\nabla_a$ and ${}^0\nabla_a$ (${}^\lambda\nabla_a g_{bc}(\lambda) = 0 = {}^0\nabla_a {}^0g_{bc}$) differ when acting on cotangent vectors by a connection in the form:

[E] For simplicity, our presentation will be focused on the vacuum case *for the unperturbed equations*, but it is straightforward to generalize it to account for matter sources.

$$ {}^{\lambda}\nabla_a \omega_b = {}^{0}\nabla_a \omega_b - C^c{}_{ab}(\lambda)\omega_c. \tag{B.23}$$

If we then write $C^c{}_{ab}(\lambda)$ in terms of ${}^0\nabla_a$ and $g_{ab}(\lambda)$, we obtain:

$$C^c{}_{ab}(\lambda) = \tfrac{1}{2} g^{cd}\Big[{}^0\nabla_a g_{bd}(\lambda) + {}^0\nabla_b g_{ad}(\lambda) - {}^0\nabla_d g_{ab}(\lambda)\Big]. \tag{B.24}$$

Then, with a little algebraic effort, we can compute the Riemann curvature tensor:

$$R_{abc}{}^d(\lambda) = {}^0R_{abc}{}^d + 2\,{}^0\nabla_{[a} C^d_{\ b]c} - 2C^e_{\ c[a} C^d_{\ b]e}, \tag{B.25}$$

from which we obtain the Ricci:

$$R_{ac}(\lambda) = {}^0R_{ac} + 2\,{}^0\nabla_{[a} C^b_{\ b]c} - 2C^e_{\ c[a} C^b_{\ b]e}. \tag{B.26}$$

Then, differentiating this entire expression with respect to λ and evaluating it at $\lambda = 0$, we obtain:

$$\dot{R}_{ac} = 2\,{}^0\nabla_{[a} \dot{C}^b_{\ b]c}, \tag{B.27}$$

where we are using a dot to denote a derivative with respect to λ at $\lambda = 0$. We note that the term quadratic in $C^c{}_{ab}$ will not yield any contribution, since $C^c{}_{ab}(\lambda{=}0) = 0$.

Thus, denoting the metric derivatives by γ, $\gamma_{ab} \equiv \dot{g}_{ab}$, and by noting that ${}^0\nabla_a\, {}^0g_{bc} = 0$, we can easily compute $\dot{C}^c{}_{ab}$, yielding:

$$\dot{C}^c{}_{ab} = \frac{1}{2}\,{}^0g^{cd}\Big[{}^0\nabla_a \gamma_{bd} + {}^0\nabla_b \gamma_{ad} - {}^0\nabla_d \gamma_{ab}\Big]. \tag{B.28}$$

Then, subtistituting on (B.27), we obtain:

$$\dot{R}_{ac} = \frac{1}{2}\,{}^0g^{bd}\Big[{}^0\nabla_a\, {}^0\nabla_c \gamma_{bd} + {}^0\nabla_b\, {}^0\nabla_d \gamma_{ac} - 2\,{}^0\nabla_b\, {}^0\nabla_{(c}\gamma_{a)d}\Big]. \tag{B.29}$$

At this point, we simplify our notation, dropping the prescript 0 for background quantities and using the background metric to raise and lower indices. In this notation, we obtain:

$$\dot{R}_{ac} = \tfrac{1}{2}\nabla_a\nabla_c\gamma + \tfrac{1}{2}\nabla^b\nabla_b\gamma_{ac} - \nabla^b\nabla_{(c}\gamma_{a)b}, \tag{B.30}$$

where we have defined $\gamma \equiv g^{ab}\gamma_{ab}$.

Now, if we multiply this entire equation by an infinitesimal variation λ, we obtain the form for infinitesimal variations of the curvature, in the familiar notation of chapter 3:

$$\delta R_{ac} = \tfrac{1}{2}g^{bd}\nabla_a\nabla_c\delta g_{bd} + \tfrac{1}{2}\nabla^b\nabla_b\delta g_{ac} - \nabla^b\nabla_{(c}\delta g_{a)b}. \tag{B.31}$$

As a final remark, we note that variations of the inverse metric g^{ab} are *not* simply given by raising the indexes of δg_{ab} with the unperturbed metric. We can calculate these by using that $g^{ab}g_{bc} = \delta^a_{\ c}$, and thus:

$$0 = \delta(g^{ab}g_{bc}) = g_{bc}\delta g^{ab} + g^{ab}\delta g_{bc} \qquad \Rightarrow \qquad \delta g_{ab} = -g_{ad}g_{bc}\delta g^{dc}. \tag{B.32}$$

B.3 Lie derivatives, Killing fields, and conformal transformations

We have already seen that a crucial operational toolbox to handle curved spaces (in the form of smooth manifolds $(\mathcal{M}, g_{ab})$) is their differential structure. We have throughout been using tangent spaces at each point to define tensor fields and compute many local quantities in our theory. Correspondingly, when one must handle *extensive* geometrical quantities and operations (such as finite arclengths or displacements, the parallel transport of vectors and tensors, etc.) one must develop a suitable integral structure to extend differential structures throughout spacetime.

In doing so, a central geometrical structure are *vector fields*, which we can used to build integral orbits (curves) and meaningfully transport local quantities throughout spacetime. Let us consider a differentiable vector field ξ^a defined on the spaces tangent to $\mathcal{M}$ at each event; we can use this field to find integral curves $C \subset \mathcal{M}$ (such that ξ^a will be tangent to them at each event), and we can define a 1-parameter family of diffeomorphisms $\phi : \mathbb{R} \times \mathcal{M} \longrightarrow \mathcal{M}$ which act translating all points along these curves by a variable amount. More precisely, for any parameters $t, s \in \mathbb{R}$, we will have diffeomorphisms obeying:

$$\phi_{t+s} = \phi_t \circ \phi_s, \qquad \Rightarrow \qquad \phi_0 = \mathbb{1}_{\mathcal{M}}. \tag{B.33}$$

These diffeomorphisms will induce natural maps between tangent vectors (or, more generally, tensors) at point p and tangent vectors (tensors) at points $\phi_t(p)$ along its orbits.

These maps are called *pushforwards* (as they "push tensors forward" from p to $\phi_t(p)$) and they are denoted ϕ_t^*:

$$T^{a_1\ldots a_l}_{b_1\ldots b_m} \in \mathcal{T}_p(l,m) \;\longrightarrow\; \phi_t^* T^{a_1\ldots a_l}_{b_1\ldots b_m} \in \mathcal{T}_{\phi_t(p)}(l,m). \tag{B.34}$$

With these maps, one can define a particular notion of a derivative to a tensor field along the integral orbits of ξ^a, the so-called Lie Derivative. Since we cannot in general subtract tensors defined at different points (*i.e.*, at different tangent spaces), a way to evaluate their difference at the points p and $\phi_t(p)$ is to "pull the latter back to p" by the induced map ϕ_{-t}^*. The Lie derivative is thus defined as:

$$\pounds_\xi (T^{a_1\ldots a_l}_{b_1\ldots b_m})_p = \lim_{t\to 0}\left[\frac{(\phi_{-t}^* T^{a_1\ldots a_l}_{b_1\ldots b_m})_p - (T^{a_1\ldots a_l}_{b_1\ldots b_m})_p}{t}\right]. \tag{B.35}$$

On the first sight, it may seem that this would just yield a directional derivative along ξ^a, $\xi^c \nabla_c T^{a_1\ldots a_l}_{b_1\ldots b_n}$. Note, however, that we have not imposed any restrictions in the magnitudes or orientations of ξ^a throughout $\mathcal{M}$ (except that they should vary smoothly), such that fixed parameter displacement t may produce displacements of varied magnitudes and directions throughout spacetime[F]. Thus, a Lie derivative will generally carry information about the variations of ξ^a as well, which will manifest in the form of terms proportional to its covariant derivative.

We shall not derive here how to obtain an expression for the Lie derivative in terms of purely geometrical operations (see appendix C of (5) for a complete and pedagogical derivation), but we quote here the result, which we shall use throughout this thesis:

$$\pounds_\xi T^{a_1\ldots a_l}_{b_1\ldots b_m} = \xi^c \nabla_c T^{a_1\ldots a_l}_{b_1\ldots b_m} - \sum_{i=1}^{l} T^{a_1\ldots c\ldots a_l}_{b_1\ldots b_m} \nabla_c \xi^{a_i} + \sum_{j=1}^{m} T^{a_1\ldots a_l}_{b_1\ldots c\ldots b_m} \nabla_{b_j} \xi^c. \tag{B.36}$$

A particular important class of diffeomorphisms in $(\mathcal{M}, g_{ab})$ are isometries, that is, transformations that leave the metric field g_{ab} invariant. A vector field that generates a 1-parameter family of isometries ϕ_t^*, $\phi_t^* g_{ab} = g_{ab}$ is called a *Killing Field*. We immediately see from equation (B.35) that the Lie derivative of the metric along any Killing field vanishes. Then, eq (B.36) yields:

$$\pounds_\xi g_{ab} = \nabla_a \xi_b + \nabla_b \xi_a = 2\nabla_{(a}\xi_{b)} = 0, \tag{B.37}$$

[F] This is particularly clear when we consider, for instance, a rotation: the magnitudes of the displacement for a fixed angular variation θ will produce larger displacements at larger radii, and go towards different directions for each position.

since $\nabla_c g_{ab} = 0$.

In fact, in terms of the Lie derivative in respect to a Killing field, we may then define a precise notion of what it means for the metric to remain constant or vary throughout spacetime, since its covariant derivative is trivially null. We can say that Minkowski spacetime, for instance, has a constant metric *everywhere*, since it is maximally symmetric, and one can get from any event p into any distinct event q by following the integral orbits of a Killing field (*i.e.* by following an isometry). In fact, the same is true for any homogeneous spacetimes, such as Einstein's Static Universe, or de Sitter spaces.

Equation (B.37) is called the Killing equation, and, by solving it, one can find the generators of various isometries in one spacetime, if it has any. Another very important family of vector fields in a spacetime are the so-called *conformal* Killing fields. They obey a relation similar to (B.37), called the conformal Killing equation:

$$\pounds_\xi g_{ab} = 2\nabla_{(a}\xi_{b)} = \lambda g_{ab}. \tag{B.38}$$

That is, the metric can only vary along a conformal Killing field parallel to itself. Thus, the associated integral transformations, called *conformal transformations*[G] will at most stretch or contract the metric, but will always preserve angles. We note, however, that not all conformal transformations (and not all isometries) must belong to a continuous group generated by a (conformal) Killing field. Generally, one can write a conformal transformation ψ as a spacetime diffeomorphism, whose induced map in the metric ($\psi^* g_{ab}$) will act in the form:

$$g_{ab}(x) \longrightarrow \tilde{g}_{ab}(x) = \Omega^2(x) g_{ab}(x), \tag{B.39}$$

where $\Omega^2(x) > 0$ is a positive function of spacetime.

Conformal transformations are particularly useful in GR because they allow one to distort spacetime distances while preserving all angles. Particularly, it is obvious from (B.39) that a vector will be spacelike, timelike or null with respect to $\tilde{g}_{ab}$ if, and only if, it is respectively spacelike, timelike or null with respect to g_{ab}. This means that any two conformally related spacetimes will have the same causal structure, even though they may have widely different geometries.

For the aforementioned reasons, it can be often more convenient to carry a geometrical or dynamical analysis originally defined in one spacetime in another, conformally related one. For such, it is useful to write geometric tensors from one spacetime in terms

[G] In the literature, they are sometimes referred to as *conformal isometries*, but we avoid this term since it may be misleading.

of the other. If ∇_a and $\tilde{\nabla}_a$ are the covariant derivatives related to g_{ab} and $\tilde{g}_{ab}$, respectively, we may express their difference through a connection $C^c_{\ ab}$, defined by:

$$\tilde{\nabla}_a \omega_b = \nabla_a \omega_b - C^c_{\ ab}\omega_c, \tag{B.40}$$

such that it can be written as:

$$\begin{aligned} C^c_{\ ab} &= \tfrac{1}{2}\tilde{g}^{cd}\Big[\nabla_a \tilde{g}_{bd} + \nabla_b \tilde{g}_{ad} - \nabla_d \tilde{g}_{ab}\Big] \\ &= \Omega^{-1}\Big[2\delta^c_{(a}\nabla_{b)}\Omega - g_{ab}g^{cd}\nabla_d\Omega\Big]. \end{aligned} \tag{B.41}$$

From this connection, one can derive with some algebraic effort the values of the tensor curvatures $\tilde{R}_{abc}^{\ \ \ d}$, $\tilde{R}_{ab}$ and $\tilde{R}$ in terms of their conformally related counterparts and Ω. Particularly, we will be interested in the Ricci curvature and the curvature scale, which read:

$$\begin{aligned} \tilde{R}^b_{\ a} = {} & \Omega^{-2}\Big\{R^b_{\ a} - (n-2)g^{bc}\nabla_c\nabla_a(\ln\Omega) - \delta^b_{\ a}g^{cd}\nabla_c\nabla_d(\ln\Omega) \\ & + (n-2)g^{bc}(\nabla_c \ln\Omega)(\nabla_a \ln\Omega) - (n-2)\delta^b_{\ a}g^{cd}(\nabla_c\ln\Omega)(\nabla_d\ln\Omega)\Big\}, \end{aligned} \tag{B.42}$$

$$\tilde{R} \ = \ \Omega^{-2}\Big\{R - 2(n-1)g^{cd}\nabla_c\nabla_d\ln\Omega - (n-1)(n-2)g^{cd}(\nabla_c\ln\Omega)(\nabla_d\ln\Omega)\Big\}. \tag{B.43}$$

Finally, we would like to analyze how our field equations transform upon a conformal transformation. Since we are often interested in working in simpler, conformally related spacetimes than that of direct interest to our problem, it would be particularly convenient to know if there is some conformal scaling that we may perform in our field, $\phi \to \tilde{\phi} = \Omega^s\phi$ (where s is called a conformal weight), so that the *form* of the field equations remains invariant. More precisely, if our field equations are defined by spacetime differential operator L_x, we would like to find an invariance in the form:

$$L_x\phi(x) = 0 \quad \Leftrightarrow \quad \tilde{L}_x\tilde{\phi}(x) = 0, \tag{B.44}$$

where $\tilde{L}_x$ is the conformally transformed operator.

A particularly simple case is that of massless scalar field for which $L = \Box = g^{ab}\nabla_a\nabla_b$. We see that a conformally transformed equation would read:

$$\begin{aligned} 0 = \tilde{g}^{ab}\tilde{\nabla}_a\tilde{\nabla}_b\tilde{\phi} &= \Omega^{-2}g^{ab}\Big[\nabla_a\nabla_b(\Omega^s\phi) - C^c_{\ ab}\nabla_c(\Omega^s\phi)\Big] \\ &= \Omega^{s-2}g^{ab}\nabla_a\nabla_b\phi + (2s+n-2)\Omega^{s-3}g^{ab}\nabla_a\Omega\nabla_b\phi \\ &\quad + s\Omega^{s-3}\phi g^{ab}\nabla_a\nabla_b\Omega + s(n+s-3)\Omega^{s-4}\phi g^{ab}\nabla_a\Omega\nabla_b\Omega. \end{aligned} \tag{B.45}$$

For this equation to be made equivalent with $g^{ab}\nabla_a\nabla_b\phi = 0$, we must require that all terms except the first identically vanish. Well, we can immediately see that, for $n \neq 2$, there will be no choice of s that allows for such cancelling. However, a very convenient covariant way of modifying this field equation (which furthermore recovers the same field equation $\Box\phi = 0$ in flat spaces) is by adding a coupling with scalar curvature, making:

$$L = g^{ab}\nabla_a\nabla_b + \xi R, \qquad \xi = cte. \tag{B.46}$$

In this case, one can verify that there is indeed a special value of ξ (which will depend on n) that will produce a convenient cancellation of terms in the modified equation, namely:

$$\xi(n) = \frac{n-2}{4(n-1)}. \tag{B.47}$$

With this value, we can put our equations in a conformally invariant form by choosing the appropriate conformal weight $s = \frac{2-n}{2}$:

$$\left[\Box + \frac{n-2}{4(n-1)}R(x)\right]\phi(x) \to \left[\tilde{\Box} + \frac{n-2}{4(n-1)}\tilde{R}(x)\right]\tilde{\phi}(x) = \Omega^{s-2}(x)\left[\Box + \frac{n-2}{4(n-1)}R(x)\right]\phi(x). \tag{B.48}$$

Particularly, for $n = 4$, we have $\xi(4) = 1/6$ and $s = -1$.

Finally, we note that we cannot maintain conformal invariance if we add a mass term to our equation (unlike the other two terms it will scale as Ω^s, not Ω^{s-2}). One can interpret this fact by noting that m will introduce a natural (inverse) length scale to the theory; since we keep this scale unchanged when we operate a conformal transformation (which distorts distances), this will necessarily provoke a nontrivial distortion in relative scales in our theory.